

Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи

Храмеева Екатерина Евгеньевна

Дальние взаимодействия в геномах эукариот и регуляция сплайсинга

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата биологических наук

Научные руководители:

д. б. н., к. ф.-м. н., профессор

Миронов А.А.

д. б. н., к. ф.-м. н., профессор

Гельфанд М.С.

Москва – 2014

Оглавление

Введение	3
Обзор литературы	11
1. Общие механизмы сплайсинга	11
2. Вторичная структура РНК и сплайсинг	18
3. Регуляция сплайсинга с помощью белковых факторов	24
4. Транс-сплайсинг	32
5. Выводы	39
Глава 1. Поиск вторичных структур РНК, участвующих в регуляции сплайсинга	41
1.1. Материалы и методы	41
1.2. Результаты и обсуждение	47
1.3. Выводы к Главе 1	55
Глава 2. Регуляция сплайсинга с помощью белковых факторов	57
2.1. Материалы и методы	57
2.2. Результаты и обсуждение	62
2.3. Выводы к Главе 2	69
Глава 3. Транс-сплайсинг	71
3.1. Материалы и методы	71
3.2. Результаты и обсуждение	75
3.3. Выводы к Главе 3	95
Заключение	96
Список литературы	100

Введение

Актуальность работы. Ни одна из естественных наук в настоящее время не обходится без применения компьютерных методов. Они позволяют хранить и обрабатывать большие объемы данных, моделировать природные процессы и системы, предсказывать их поведение. Так, на стыке биологии и компьютерных наук появилось новое самостоятельное научное направление — биоинформатика, которая использует компьютерные средства для решения биологических задач.

По мере развития экспериментальных методов, секвенирование геномов становится все более быстрым и дешевым процессом. Темпы секвенирования значительно опережают темпы экспериментального анализа геномов, и изучение структуры и функции ДНК, РНК и белков на всех этапах включает использование специальных вычислительных методов. Наличие большого количества расшифрованных геномов позволяет предсказывать функции генов и механизмы их регуляции в новых геномах по аналогии с уже изученными геномами. Задача изучения регуляции генов особенно интересна для многоклеточных эукариот, так как их гены имеют наиболее сложную структуру и считываемая с них пре-мРНК часто альтернативно сплайсируется.

Альтернативный сплайсинг — один из основных механизмов создания разнообразия белковых последовательностей. Альтернативный сплайсинг может вносить незначительные изменения в структуру и функцию белка, может резко изменять их, может приводить к формированию нетранслируемых изоформ. Альтернативный сплайсинг не только является одним из важных механизмов регуляции функционального состояния белков (а значит клеток и организмов в целом), но и сам подвержен сложной регуляции.

Известно, что у многих организмов вторичная структура РНК оказывает значительное влияние на процессы транскрипции и трансляции, однако влияние

таких структур на процесс сплайсинга до сих пор систематически не изучено. Этот вопрос интересовал исследователей ещё со времен открытия сплайсинга, однако считалось, что основную роль в регуляции сплайсинга играют транс-факторы (элементы сплайсосомы, отдельные вспомогательные белки, различные регуляторные РНК). Недавние исследования показали, что механизмы сплайсинга, опосредованные вторичной структурой, могут быть более распространёнными, чем предполагалось ранее [1]. Поэтому возникла необходимость систематического исследования окружений донорных и акцепторных сайтов сплайсинга на предмет наличия консервативных вторичных структур РНК, которые могли выступать в качестве регуляторов сплайсинга.

Вторичная структура РНК не является единственным механизмом регуляции альтернативного сплайсинга. Белковые факторы также способны оказывать влияние на сплайсинг, связываясь с особыми регуляторными последовательностями — энхансерами и сайленсерами. Энхансеры способствуют вырезанию интрона, а сайленсеры, напротив, противодействуют. Известно несколько примеров регуляции сплайсинга белком hnRNPL, принадлежащим к семейству белков hnRNP — высоко экспрессированных в клетке белков, выполняющих разнообразные функции в метаболизме РНК. Функция белка hnRNPL до конца не ясна, поэтому, учитывая его широкую представленность в клетке, представляется интересным изучить механизмы регуляции альтернативного сплайсинга белком hnRNPL на полногеномном уровне.

Транс-сплайсинг — это особая форма процессинга РНК, в результате которой экзоны, находящиеся на двух разных молекулах РНК, соединяются и лигируются. Секвенирование транскриптомов различных организмов, от червей до человека, показало, что многие транскрипты состоят из сегментов последовательностей, которые не следуют друг за другом на хромосоме, а происходят из удалённых участков генома, и даже с разных хромосом. Некоторые из таких химерных транскриптов образуются в результате генетических перестроек

у объекта секвенирования (чаще всего, в опухолевых клетках). Другие возникают в нормальных тканях пост-транскрипционно в результате транс-сплайсинга. Это указывает на то, что транс-сплайсинг у млекопитающих распространен существенно шире, чем считалось ранее, что делает задачу его изучения крайне важной.

Цель диссертационной работы состоит в изучении механизмов регуляции альтернативного сплайсинга вторичными структурами РНК и белковыми факторами, а также распространенности транс-сплайсинга, на полногеномном уровне.

Для достижения поставленной цели были решены следующие задачи:

1. Разработать метод поиска консервативных вторичных структур, ассоциированных со сплайсингом.
2. Исследовать окружения донорных и акцепторных сайтов сплайсинга на предмет наличия вторичных структур РНК, которые могли бы регулировать сплайсинг.
3. Изучить полногеномную карту позиций связывания белка hnRNPL с РНК в клетках человека HeLa.
4. Установить взаимосвязь между относительной позицией контакта белка hnRNPL с РНК и его активаторным или репрессирующим влиянием на сплайсинг.
5. Осуществить поиск химерных транскриптов, содержащих последовательности, соответствующие разным хромосомам.
6. Проверить, производят ли пространственно близкие участки генома большое количество химерных РНК по механизму транс-сплайсинга.

Научная новизна и практическая значимость. Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, — с помощью хэширования. Этот метод является альтернативой широко применяемому методу построения вторичных структур РНК с использованием множественного выравнивания [2]. Преимуществом метода хэширования является вычислительная быстрота, позволяющая осуществлять поиск вторичных структур РНК в полногеномном масштабе для многоклеточных эукариот без использования выравнивания. Применение множественного выравнивания ко вторичным структурам, определенным с помощью хэширования, позволяет уточнять предсказания, незначительно увеличивая время расчета.

В настоящее время экспериментально показано наличие около 15 случаев участия вторичной структуры в регуляции сплайсинга у различных организмов [3]: у вирусов (вирус гепатита В, аденовирус, вирус иммунодефицита человека типа 1, вирус саркомы Рауса), дрожжей, растений (*Nicotiana plumbaginifolia*), насекомых (*Drosophila*), а также у крыс и мышей. Известно, что в организме человека вторичные структуры могут оказывать влияние на эффективность распознавания сайтов сплайсинга и таким образом участвовать в формировании изоформ гормона роста, гена *tau*, гена *Hprt* и гена *hnRNPA1*. Ошибки сплайсинга, обусловленные влиянием вторичных структур РНК, приводят к таким патологиям человека, как мышечная дистрофия, кистозный фиброз и паркинсонизм. В настоящей работе идентифицировано несколько сотен генов, содержащих потенциальные консервативные вторичные структуры. Это наблюдение позволяет предположить, что механизмы сплайсинга, опосредованные вторичной структурой, могут быть более распространенными, чем предполагалось ранее.

Впервые получена и проанализирована точная полногеномная карта позиций связывания белка hnRNPL с РНК в клетках человека HeLa. Известно несколько случаев регуляции сплайсинга с помощью белка hnRNPL. Белок

hnRNPL регулирует пропуск экзона в гене *CD45* в ответ на активацию Т-клеток. Он также оказывает влияние на альтернативный сплайсинг других генов посредством удерживания интрона, подавления включения нескольких экзонов и выбора альтернативного сайта полиаденилирования. Белок hnRNPL взаимодействует с 3' нетранслируемой областью (3'НТО) мРНК синтазы оксида азота и регулирует её стабильность. Белок hnRNPL принадлежит обширному семейству белков hnRNP, которые являются одними из наиболее высоко экспрессируемых в клетке и выполняют разнообразные функции в метаболизме пре-мРНК, среди которых упаковка только что синтезированных транскриптов, регуляция конститутивного и альтернативного сплайсинга, транспорт молекул мРНК и их локальная трансляция, регуляция стабильности мРНК, активация или репрессия трансляции. В настоящей работе показано, что участие белка hnRNPL в регуляции сплайсинга не ограничивается отдельными случаями, а носит полногеномный характер.

Транс-сплайсинг встречается не только у трипаносом и нематод, как считалось ранее. Ранее были экспериментально показаны два случая транс-сплайсинга в клетках человека. Транс-сплайсинг может происходить между 5' экзонами гена JAZF1 на локусе 7p15 и 3' экзонами гена JJAZ1 на локусе 17q11, причем получившийся в результате химерный транскрипт транслируется в белок, препятствующий апоптозу. Транс-сплайсинг наблюдается и между генами SLC45A3 и ELK4, также с образованием функционального белка. Кроме того, данные высокопроизводительного секвенирования указывают на то, что транс-сплайсинг у млекопитающих — не редкое явление, как считалось ранее, а довольно распространенный механизм, что подтверждается результатами настоящей работы.

Впервые систематически изучено функциональное состояние часто контактирующих участков ДНК, находящихся на разных хромосомах. С помощью полногеномной карты частот контактов участков ДНК показано, что часто контактирующие фрагменты имеют сходный уровень экспрессии, модификаций хро-

матина, метилирования ДНК, чувствительности к ДНКазе, а также производят большое количество химерных РНК, большая часть которых, по-видимо, имеет пост-транскрипционное происхождение и образуется в результате транс-сплайсинга. Это наблюдение подтверждает существование транскрипционных фабрик, обогащенных факторами транскрипции и сплайсинга, в которых активно экспрессирующиеся ко-регулируемые гены могут образовывать межхромосомные контакты. Возможность организации генов в транскрипционные фабрики открывает новый, более сложный уровень регуляции генной активности и показывает, что современные представления о многокомпонентной системе регуляции экспрессии генов у многоклеточных эукариот являются лишь вершиной айсберга.

На защиту выносятся следующие основные результаты и положения. Разработан метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, основанный на хэшировании. У млекопитающих идентифицировано несколько сотен генов, содержащих потенциальные консервативные вторичные структуры. Разработан метод оценки уровня ложных положительных предсказаний и показано, что по пессимистичной оценке этот уровень не превосходит 30%. Консервативные вторичные структуры часто встречаются в альтернативно сплайсируемых генах и могут осуществлять регуляцию альтернативного сплайсинга. В частности, образование вторичной структуры необходимо для правильного вырезания интрона в гене *SF1*.

Изучена полногеномная карта позиций связывания белка hnRNPL с РНК в клетках человека HeLa. Сайты связывания обогащены CA-повторами и CA-богатыми мотивами. Распределение позиций связывания hnRNPL вокруг 5'- и 3'-сайтов сплайсинга различается между альтернативными и константными экзонами, и между L-активируемыми и L-репрессируемыми экзонами. Позиция связывания белка hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг. Белок hnRNPL часто связывается вблизи мишеней мик-

роРНК в области 3'UTR. Плотность позиций связывания hnRNPL в интронах, содержащих snoRNA, значительно выше, что может говорить об участии белка hnRNPL в синтезе snoRNA.

В данных о пространственной близости фрагментов генома человека обнаружены и удалены систематические ошибки секвенирования. Анализ систематических ошибок секвенирования показал зависимость профилей покрытия генов от лаборатории в экспериментах по секвенированию мРНК. В трех наборах данных секвенирования транскриптома человека (ткань мозга, клеточная линия эритролейкемии K562, лимфобластоидная клеточная линия GM12878) найдены химерные РНК. Пространственно близкие фрагменты ДНК образуют между собой больше химерных РНК, чем пространственно далекие, по-видимому, в основном за счет транс-сплайсинга. Пространственно близкие фрагменты ДНК характеризуются схожими эпигенетическими маркерами и состояниями хроматина, гены в них функционально подобны и ко-экспрессируются, что хорошо согласуется с теорией о фабриках транскрипции.

Апробация работы Материалы исследований по теме диссертации были представлены на международных конференциях: XV Международной конференции студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2008, диплом за лучший доклад), I Международном конкурсе научных работ молодых ученых в области нанотехнологии "Роснанотех" (Москва, 2008, призер конкурса), XVII Международной конференции студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2010), Bioinformatics after Next Generation Sequencing (Звенигород, 2010), 18th annual international conference on Intelligent Systems for Molecular Biology ISMB (Бостон, 2010), 24th International Mammalian Genome Conference (Крит, 2010), Albany 2011: The 17th Conversation (Олбани, 2011), Moscow Conference on Computational Molecular Biology MCCMB (Москва, 2011), 1st Cold Spring Harbor Asia conference on Bioinformatics of Human and Animal Genomes (Сучжоу, 2011), 16th Annual International Conference on Research

in Computational Molecular Biology RECOMB (Барселона, 2012), Chromosomes, Stem Cells and Disease (Барселона, 2012), а также на конференциях ”Информационные технологии и системы” ИТиС-31 (Геленджик, 2008), ИТиС-33 (Геленджик, 2010), ИТиС-34 (Геленджик, 2011), ИТиС-35 (Петрозаводск, 2012).

Публикации Материалы диссертации опубликованы в 19 печатных работах, из них 4 статьи в рецензируемых журналах и 15 тезисов докладов.

Личный вклад автора Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Подготовка к публикации полученных результатов проводилась совместно с соавторами, причем вклад диссертанта был определяющим. Все представленные в диссертации результаты получены лично автором.

Структура и объем диссертации Диссертация состоит из введения, обзора литературы, 3 глав, заключения и библиографии. Общий объем диссертации 108 страниц, из них 97 страницы текста, включая 42 рисунка и 3 таблицы. Библиография включает 87 наименований на 9 страницах.

Обзор литературы

1. Общие механизмы сплайсинга

В 1977 году было впервые показано, что гены высших организмов содержат чередующиеся последовательности двух типов – экзоны и интроны. Экзоны впоследствии входят в состав зрелых мРНК, а интроны не входят. Экзоны у многоклеточных организмов, как правило, намного короче интронов, и лишь небольшая часть нуклеотидной последовательности гена кодирует белок. Как экзоны, так и интроны транскрибируются в РНК. Однако интроны впоследствии удаляются из синтезированных РНК в ходе процесса, называемого сплайсингом (Рисунок 1А).

Подавляющая доля сплайсинга транскриптов, происходящего в клетке, приводит к формированию кодирующих белок мРНК, и в этой работе рассматривается именно такой вид сплайсинга. Его называют сплайсингом предшественников мРНК, или сплайсингом пре-мРНК. Каждое событие сплайсинга приводит к вырезанию одного интрона; при этом происходит две последовательных реакции трансэтерификации, которые соединяют два экзона, а интрон при этом удаляется (Рисунок 1Б).

Удаление большого числа интронов в результате сплайсинга может показаться нерациональным, но этот процесс имеет преимущества. Во-первых, наличие экзон-интронной структуры способствует появлению новых полезных белков в ходе эволюции, т.к. в результате рекомбинации экзоны разных генов объединяются, и новые белки могут собираться из частей уже существующих белков в различных комбинациях. Это утверждение поддерживается тем фактом, что многие белки состоят из доменов – аминокислотных последовательностей, которые встречаются в совершенно разных белках и способны функционировать отдельно от остального белка, и во многих случаях экзоны кодируют

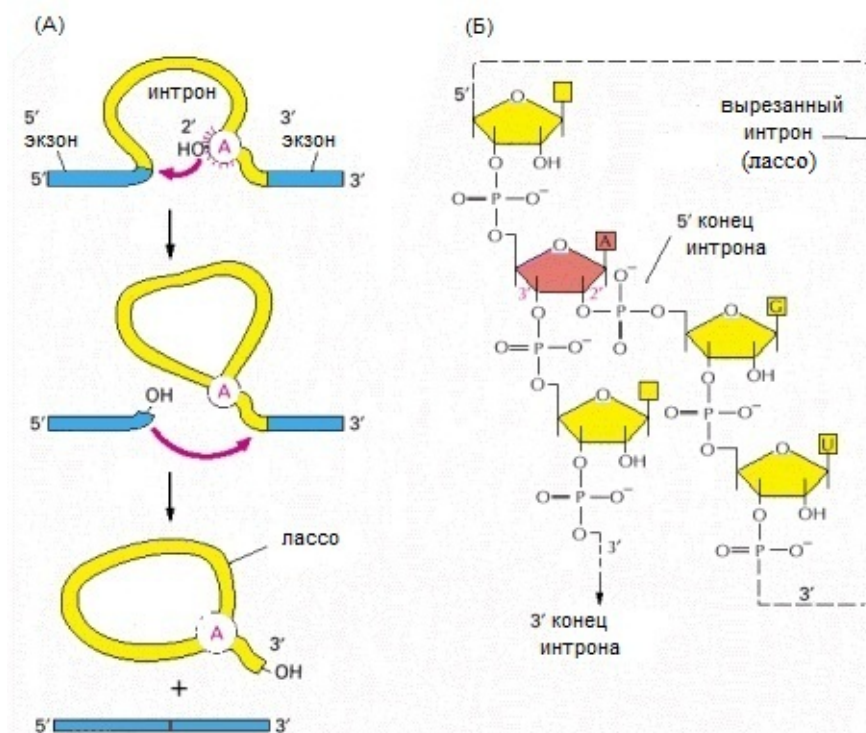


Рис. 1. Реакция сплайсинга РНК. (А) На первом шаге, 2'-ОН особого аденина (сайт ветвления, или 'branch point', выделен красным) в последовательности интрона взаимодействует с 5'-сайтом сплайсинга и разрывает сахарофосфатную связь в данной точке. При этом 5'-конец интрона ковалентно связывается с аденином, как показано справа (Б), и формируется структура в форме лассо. На втором шаге, образовавшийся 3'-ОН на конце первого экзона взаимодействует с началом второго экзона. В результате между экзонами формируется ковалентная связь, а интрон высвобождается (по [4]).

целые домены [5]. Во-вторых, транскрипты многих эукариотических генов (до 75% генов человека [6]) могут сплайсироваться большим числом различных способов; в результате один ген кодирует не один, а множество белков (Рисунок 2).

1.1. Альтернативный сплайсинг

Процесс, в результате которого первичный транскрипт может сплайсироваться разными способами и давать начало разным мРНК, называется альтернативным сплайсингом. Например, ген *DSCAM*, который отвечает за развитие

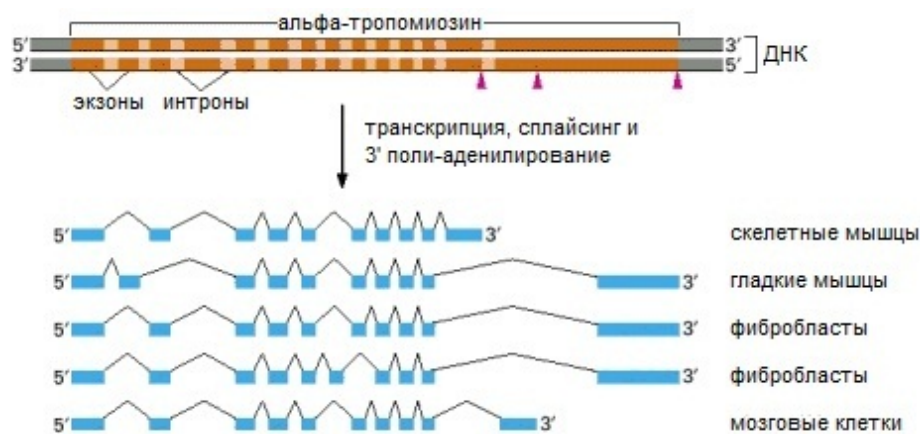


Рис. 2. Альтернативный сплайсинг гена α -тропомиозина крысы. α -тропомиозин – белок, регулирующий сокращение в мышечных клетках. Первичный транскрипт может сплайсироваться разными способами, как показано на рисунке, и образовывать разные мРНК, которые дают начало разным белкам. При этом важно, что некоторые мРНК специфичны для определенных типов тканей (по [4]).

нервной системы у представителей рода *Drosophila*, потенциально кодирует 38 тысяч разных белков [7]. Каждая зрелая молекула РНК содержит 1 из 12 альтернатив экзона А, 1 из 48 альтернатив экзона В, 1 из 33 альтернатив экзона С и 1 из 2 альтернатив экзона D. Таким образом, число возможных комбинаций составляет 38,016.

Альтернативный сплайсинг часто вызывает изменения в аминокислотной последовательности белка, определяет расположение сайта полиаденилирования, приводит к сдвигу рамки считывания и появлению стоп-кодонов внутри последовательности. Выделяют несколько основных механизмов альтернативного сплайсинга: кассетные экзоны, взаимоисключающие экзоны, удержанные интроны, альтернативные 5'- и 3'-сайты сплайсинга и др. (Рисунок 3).

1.2. Сигналы сплайсинга в нуклеотидных последовательностях

Из механизма сплайсинга, приведенного на рис. 1, видно, что в процессе удаления интрона принимают участие три позиции на РНК: 5'-сайт сплайсинга, 3'-сайт сплайсинга и сайт ветвления – особая точка в последовательности ин-

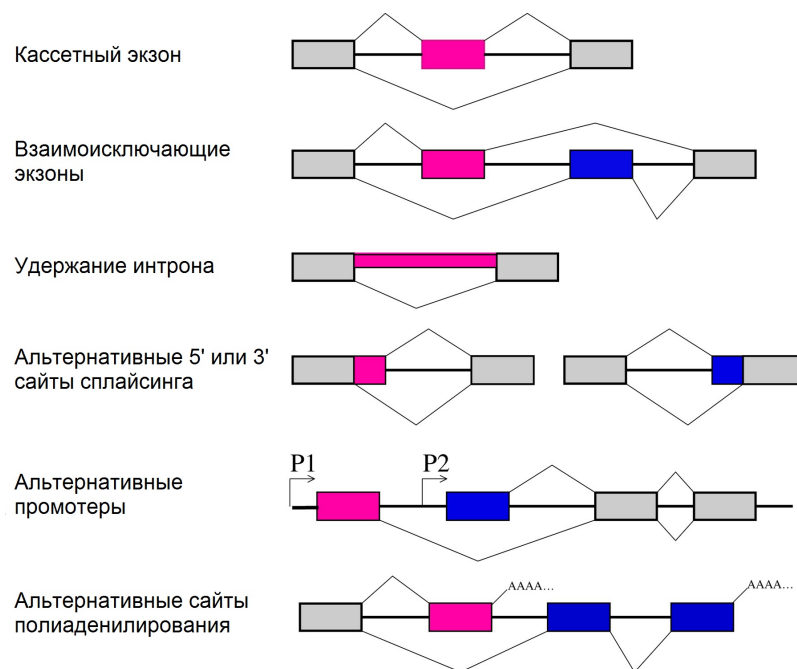


Рис. 3. Основные типы альтернативного сплайсинга. Экзоны обозначены прямоугольниками, интроны – линиями. Альтернативные экзоны выделены цветом, конститутивные экзоны – серым. Промоторы показаны стрелочками, сайты полиаденилирования - АААА (по [8]).

трона, которая формирует основу вырезаемого в виде 'лассо' интрона. Каждый из этих сайтов имеет консенсусную (похожую, но не одинаковую в интронах разных генов) последовательность, которая позволяет клеткам находить сайты сплайсинга (Рисунок 4). Возможность осуществлять сплайсинг при вариабельности сайтов крайне важна для клетки. Такая пластичность позволяет клеткам регулировать образование разных мРНК из одного первичного транскрипта, делая возможным протекание альтернативного сплайсинга и, в итоге, синтез нескольких белковых продуктов на основе одного гена.

Пластичность сплайсинга также обеспечивается наличием регуляторных сайтов - энхансеров и сайленсеров, которые, как правило, располагаются вблизи сайтов сплайсинга и регулируют эффективность вырезания интрона. Энхансеры способствуют вырезанию интрона посредством связывания белков-активаторов сплайсинга. Сайленсеры, напротив, противодействуют вырезанию интрона, связывая белки-репрессоры, подавляющие сплайсинг.

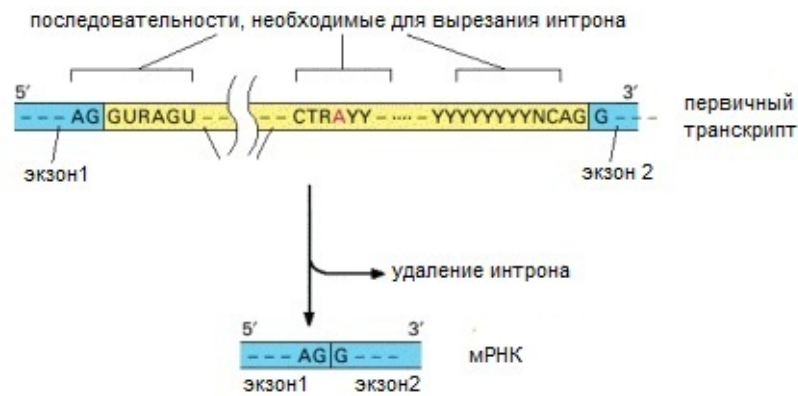


Рис. 4. Сигналы сплайсинга в интронах человека. Для сплайсинга необходимы три нуклеотидных блока, изображенных на рисунке. Остальные нуклеотиды могут быть любыми. R обозначает либо А, либо G; Y обозначает либо С, либо U. Аденин А, выделенный красным, представляет собой сайт ветвления. Динуклеотид GU в начале интрона и AG в его конце, как правило, инвариантны. Остальные нуклеотиды могут меняться, хотя изображенные варианты более предпочтительны (по [4]).

1.3. Сплайсосома – 'машина', осуществляющая сплайсинг

В осуществлении сплайсинга пре-мРНК принимают участие не только отдельные белки, но и сложные рибонуклеопротеидные комплексы. Они распознают границы интронов и катализируют химические превращения в ходе сплайсинга. В состав этих комплексов входят относительно короткие молекулы РНК (менее 200 нуклеотидов каждая); пять из них (U1, U2, U4, U5 и U6) участвуют в осуществлении наиболее распространенной формы сплайсинга. Их называют мяРНК (малые ядерные РНК, или snRNA), и каждая из этих молекул связана в комплекс как минимум с семью белковыми субъединицами, формируя таким образом мяРНП (малый ядерный рибонуклеопротеин, или snRNP). Такие мяРНП формируют ядро сплайсосомы – большого комплекса РНК и белковых молекул, который осуществляет сплайсинг пре-мРНК в клетках.

Сплайсосома – динамическая машина. Она собирается на пре-мРНК из отдельных компонентов, которые приходят и уходят из комплекса в ходе сплайсинга (Рисунок 5). Сначала ВВР (белок, связывающийся с сайтом ветвления,

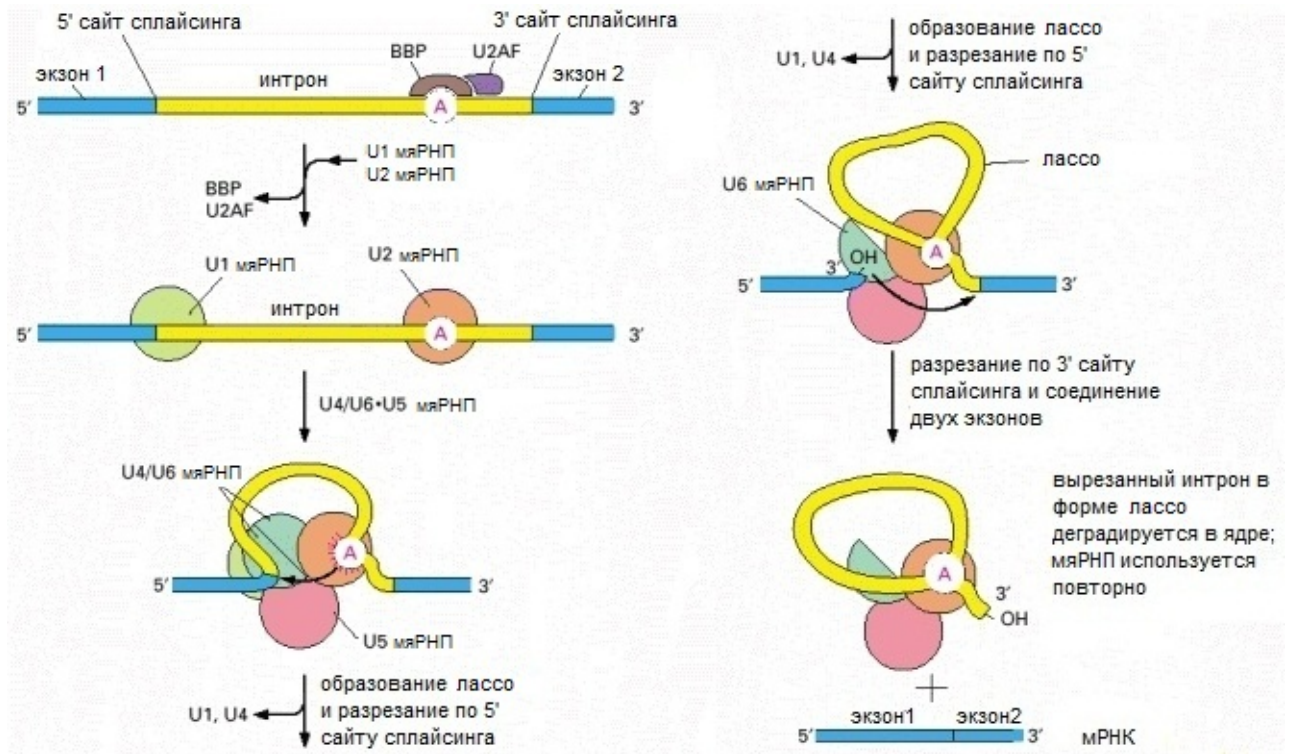


Рис. 5. Механизм сплайсинга РНК. Сплайсинг РНК катализируется мяРНП (показаны цветными кругами) и другими белками (большинство из них не показаны), которые в комплексе составляют сплайсосому. Сплайсосома узнает сигналы сплайсинга на молекуле пре-мРНК, стягивает вместе два конца интрона и обеспечивает ферментативный катализ двух реакций сплайсинга (по [4]).

или branch-point binding protein) и U2AF (вспомогательный белок) узнают сайт ветвления. На следующем шаге мяРНП U2 вытесняет BVP и U2AF, и его мяРНК формирует комплементарные пары с последовательностью сайта ветвления, а мяРНК U1 формирует комплементарные пары с 5' сайтом сплайсинга (Рисунок 4). В этот момент комплекс из трех мяРНП U4/U6-U5 присоединяется к сплайсосоме. В этом комплексе мяРНП U4 и U6 прочно связаны друг с другом посредством взаимодействий комплементарных пар в соответствующих мяРНК. Затем происходит перегруппировка мяРНК, входящих в состав комплекса, в результате которых комплементарные взаимодействия U4/U6 разрушаются, U4 мяРНП выбрасывается из сплайсосомы, а U6 мяРНП занимает место U1 на 5' сайте сплайсинга. Дальнейшие перестановки приводят к фор-

мированию активного сайта сплайсосомы, что делает возможным протекание двух трансэтерификационных реакций сплайсинга.

Низшие эукариоты (например, дрожжи) имеют только один набор мяРНП, который осуществляет весь сплайсинг пре-мРНК. Однако более сложные эукариоты, такие как насекомые, млекопитающие и растения, имеют второй набор мяРНП, который управляет сплайсингом небольшой части интронов. Эта так называемая минорная форма сплайсосомы распознает другие последовательности 5'-сайта сплайсинга, 3'-сайта сплайсинга и сайта ветвления. Ее также называют АТ-АС сплайсосомой из-за нуклеотидных детерминант границ интрона, а мажорную форму сплайсосомы, по аналогии, GT-AG сплайсосомой. В остальном минорная форма сплайсосомы мало отличается от мажорной формы (Рисунок 6).

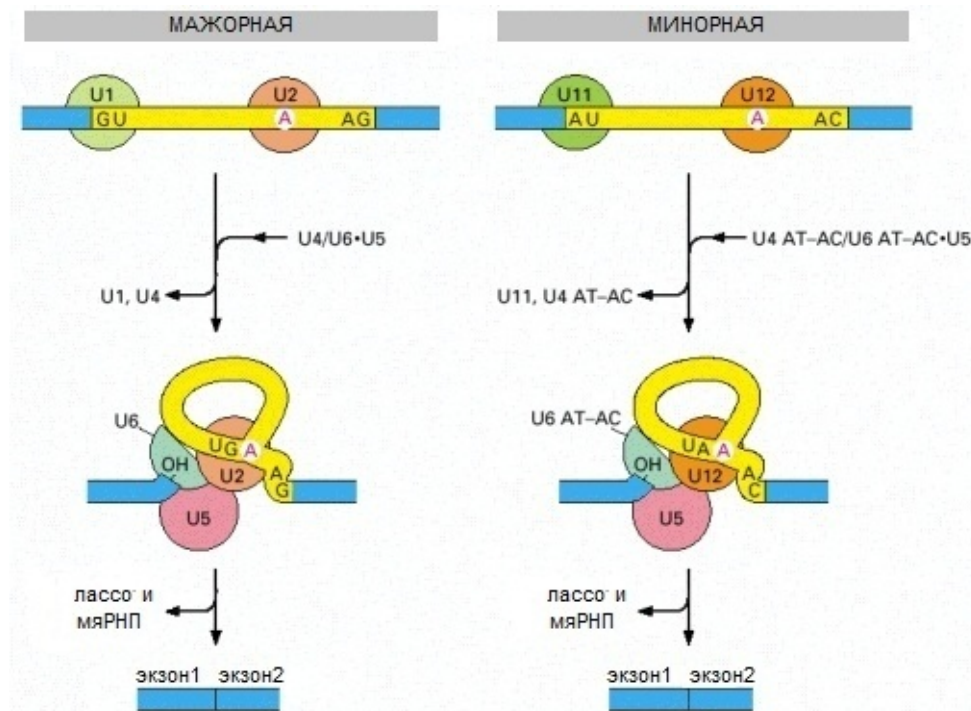


Рис. 6. Сравнение мажорной и минорной сплайсосом. Мажорная сплайсосома слева, АТ-АС сплайсосома справа. U5 snRNP – единственный общий для двух сплайсосом компонент. У человека порядка 0.1% интронов удаляется минорной сплайсосомой (по [4]).

2. Вторичная структура РНК и сплайсинг

Большое количество факторов, входящих в состав сплайсосомы, необходимо потому, что на процесс распознавания границ экзона оказывают влияние многочисленные особенности пре-мРНК, такие как длина экзона, наличие энхансеров или сайленсеров, сила сигналов сплайсинга и архитектура промотора. Также предполагается, что на активность сплайсинга может оказывать влияние вторичная структура пре-мРНК.

Влияние вторичной структуры на связывание факторов сплайсинга

Результаты нескольких исследований указывают на то, что вторичная структура РНК крайне важна для связывания регуляторных белков [3]. Так, показано, что белки семейств *muscleblind* и *CELF* способны связываться с тринуклеотидными повторами (CNG)_n на РНК, причем нарушения связывания этих белков приводят к развитию миотонической дистрофии [9]. Это наблюдение хорошо согласуется с результатами эксперимента *in vitro*, которые указывают на способность тринуклеотидных повторов (CNG)_n сворачиваться в петли характерной формы [9].

Известно, что белок *HuR*, который принимает участие в посттранскрипционной регуляции экспрессии генов, при связывании распознает определенный мотив на мРНК, ассоциированный со вторичной структурой (Рисунок 7а) [10]. Кроме того, связывание некоторых активаторов (*B52*, *SRp55* и *NOVA-1*) и репрессоров (*hnRNP A1*) сплайсинга зависит не только от нуклеотидной последовательности сайтов связывания, но и от вторичной структуры РНК [11]. Не только вторичная структура может регулировать связывание факторов сплайсинга, но и факторы сплайсинга могут влиять на образование вторичной структуры, что, по всей видимости, обеспечивает более гибкую регуляцию сплайсинга. Так, связывание фактора *U2AF65* с 3'-сайтом сплайсинга приводит к такому

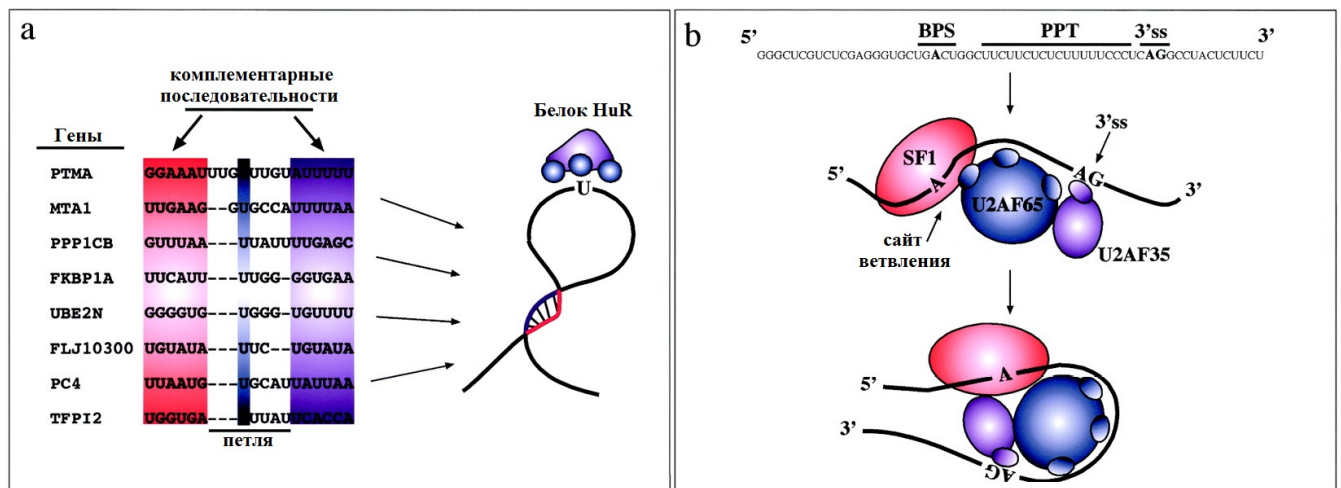


Рис. 7. Влияние вторичной структуры на связывание белков. (а) Нестрогие консервативные по последовательности сайты связывания на РНК могут представлять собой шпильки одинаковой формы, как показано для белка HuR, который связывается с AU- и U-богатыми участками мРНК и оказывает влияние на их стабильность и трансляцию. Эти участки состоят из двух стеблевых 6-нуклеотидных мотивов (выделены красным и фиолетовым) и консервативного U в петле (выделен синим). (б) Факторы сплайсинга способны влиять на пространственное расположение элементов мРНК. Так, связывание SF-1, U2AF65 и U2AF35 на 62-нуклеотидном участке РНК, содержащем сайт ветвления, полипиримидиновый тракт и 3'-сайт сплайсинга, приводит к изгибанию РНК таким образом, что сайт ветвления и 3'-сайт сплайсинга оказываются рядом; в результате облегчается формирование комплекса (по [3]).

сворачиванию мРНК, что 3'-сайт сплайсинга и сайт ветвления сближаются (Рисунок 7b) [12].

2.1. Влияние вторичной структуры на цис-элементы

Наибольший интерес при изучении влияния вторичных структур РНК на процесс сплайсинга представляют цис-элементы (3'-сайт сплайсинга, 5'-сайт сплайсинга и сайт ветвления). Предполагается, что вторичная структура может оказывать влияние на эти участки по трем различным механизмам (Рисунок 8). Наличие структурных элементов может ограничивать доступность факторов сплайсинга к сигналам сплайсинга, при этом элементы вторичной структуры могут как задерживать процесс сплайсинга, так и вызывать про-

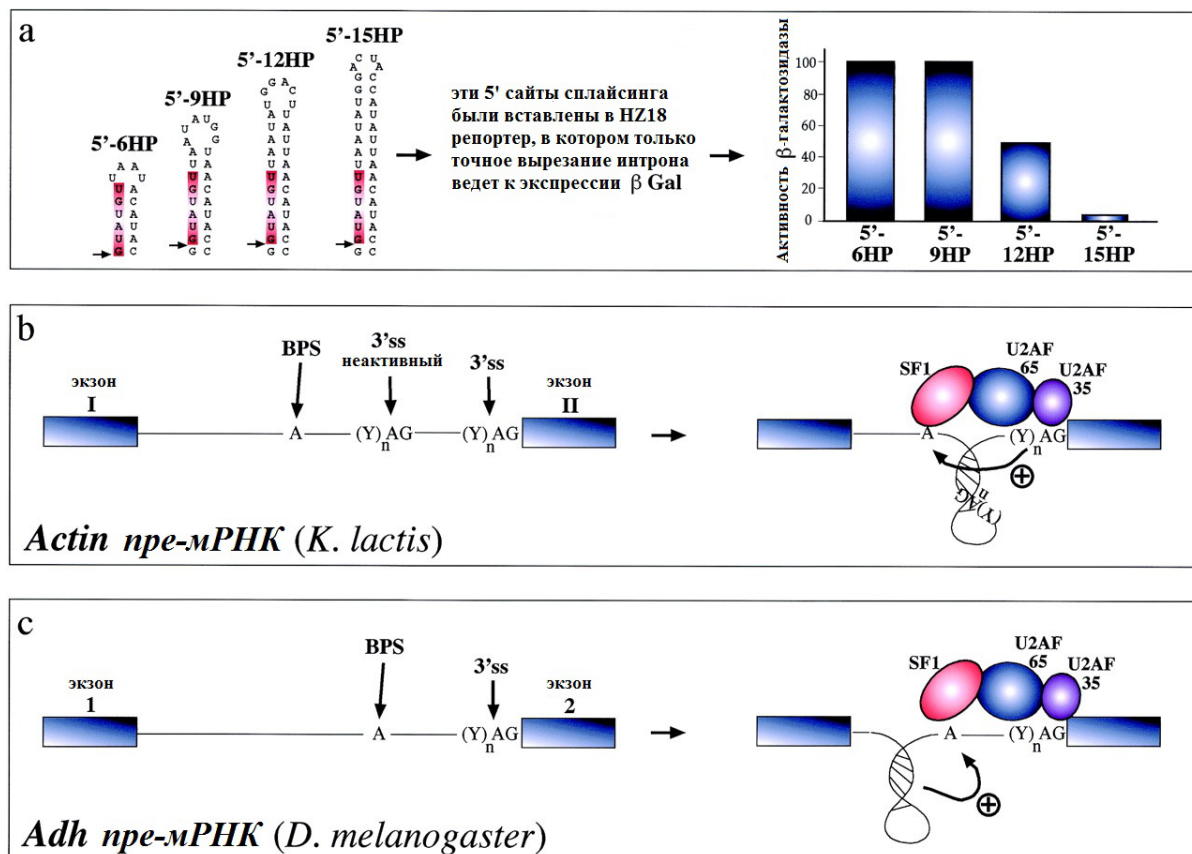


Рис. 8. Структурные элементы РНК и эффективность сплайсинга. (а) Вторичные структуры (НР) возрастающей длины, содержащие 5'-сайт сплайсинга (показан стрелочкой), встроены в ген *lacZ*. Длинные структуры изолируют сайт сплайсинга и ингибируют сборку сплайсосомы на ранних стадиях, что приводит к снижению экспрессии β -галактозы. (б) В интроне гена *Actin* дистальный 3'-сайт сплайсинга расположен слишком далеко от сайта ветвления, поэтому в вырезании интрона должен участвовать проксимальный 3'-сайт. Однако на месте проксимального 3' сайта сплайсинга образуется вторичная структура, которая сближает сайт ветвления и дистальный 3'-сайт, а также делает проксимальный 3'-сайт недоступным для факторов сплайсинга. (с) Образование вторичной структуры около сайта ветвления в первом интроне гена *Adh* поддерживает сайт ветвления в одноцепочечном состоянии, что делает его более доступным для сплайсосомы (по [3]).

пуск экзона (Рисунок 8а) [13]. Вторичные структуры могут менять расстояния между регуляторными сайтами сплайсинга, оказывая влияние на эффективность сплайсинга и частоту использования определенных сайтов сплайсинга (Рисунок 8б) [14]. Кроме того, структурные образования могут стимулировать использование сайта ветвления путем поддержания его в одноцепочечном со-

стоянии (Рисунок 8с) [15].

2.2. Влияние дальних взаимодействий во вторичной структуре на доступность участков пре-мРНК

У дрозофилы один и тот же ген *Dscam* теоретически может кодировать более 38 тысяч разных сплайс-изоформ. При этом некоторые экзоны оказываются взаимоисключающими: один и только один экзон из конкретного набора входит в зрелую мРНК. Показано, что взаимоисключающий характер сплайсинга обязан своим происхождением группе альтернативных структур РНК [16]. Перед всеми экзонами из взаимоисключающего набора располагается так называемый селекторный сайт, который комплементарен адаптерным участкам, расположенным около соответствующих экзонов. Таким образом, один и только один адаптер может комплементарно взаимодействовать с селекторным сайтом (Рисунок 9).

Влияние вторичной структуры на энхансеры и сайленсеры

В большинстве случаев экзонные и интронные энхансеры и сайленсеры регулируют сплайсинг посредством связывания различных транс-факторов (РН-Ковой или белковой природы), которые активируют или подавляют распознавание сайтов сплайсинга или сборку сплайсосомы. Необходимо заметить, что принципиальная разница между механизмами работы экзонных и интронных энхансеров и сайленсеров отсутствует. Эти регуляторные элементы до сих пор недостаточно изучены, что главным образом связано со сложностью их идентификации. Однако есть несколько экспериментальных подтверждений влияния вторичной структуры на работу энхансеров и сайленсеров.

Показано, что последовательность экзонного сайленсера человека в экзоне фибронектина *EDA* влияет на связывание регуляторных белков SR с экзонным

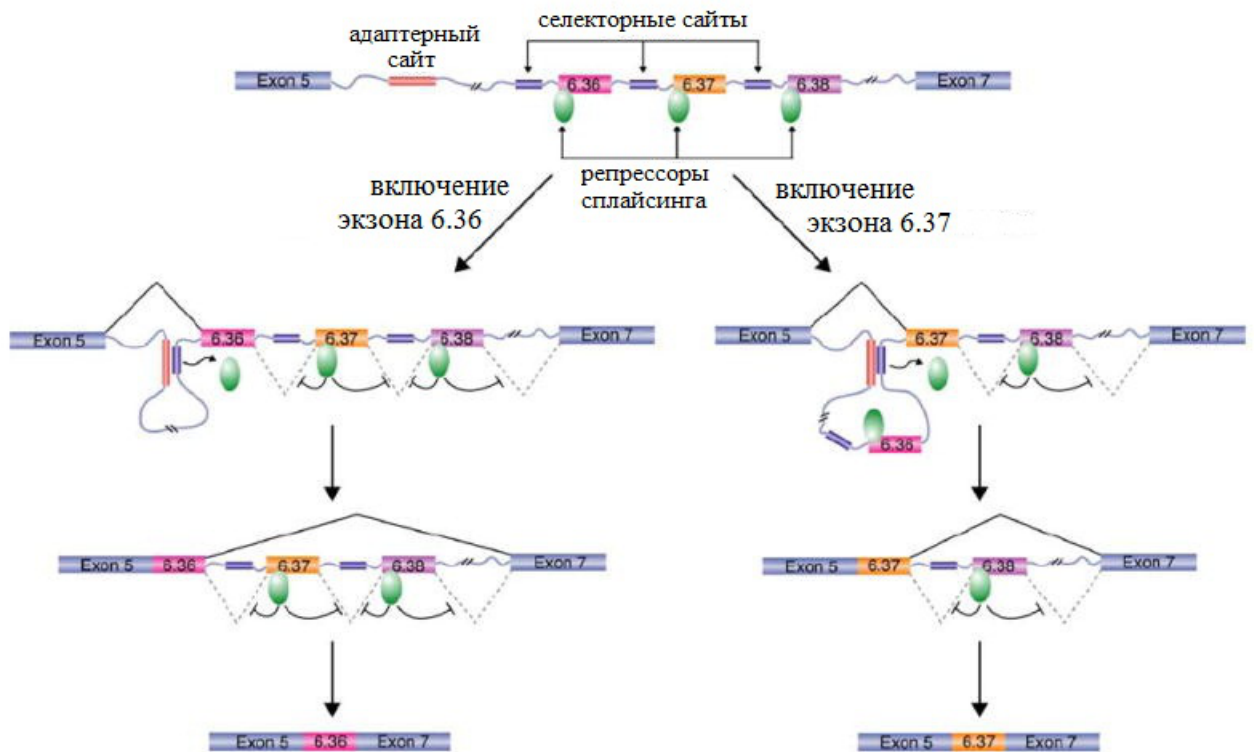


Рис. 9. Модель взаимоисключающего сплайсинга гена *Dscam*. Репрессор сплайсинга (зеленый овал) предотвращает одновременное включение нескольких вариантов экзона 6 в зрелую мРНК. Для включения варианта экзона 6 в мРНК необходимо, чтобы селекторный сайт связался с адаптерным сайтом, расположенным перед этим вариантом. Образовавшаяся в результате вторичная структура инактивирует репрессор сплайсинга по неизвестному механизму. Это приводит к включению следующего за адаптерным сайтом варианта экзона 6 в зрелую мРНК. Остальные варианты экзона 6 остаются репрессированными и не включаются в зрелую мРНК (по [16]).

энхансером [17]. Экзонный сайленсер стабилизирует вторичную структуру экзонного энхансера, способствуя связыванию белков SR. Мутации в экзонном сайленсере гена мыши не влияют на сплайсинг, а в клетках человека – влияют, что может быть объяснено сдвигом экзонного энхансера из петли в область комплементарных взаимодействий. Такой сдвиг препятствует связыванию белкового регулятора SF2/ASF и приводит к пропуску экзона (Рисунок 10а).

Кроме того, для человеческого гена *FGFR2* показано образование вторичной структуры РНК из одноцепочечных элементов IAS2 (интронная акти-

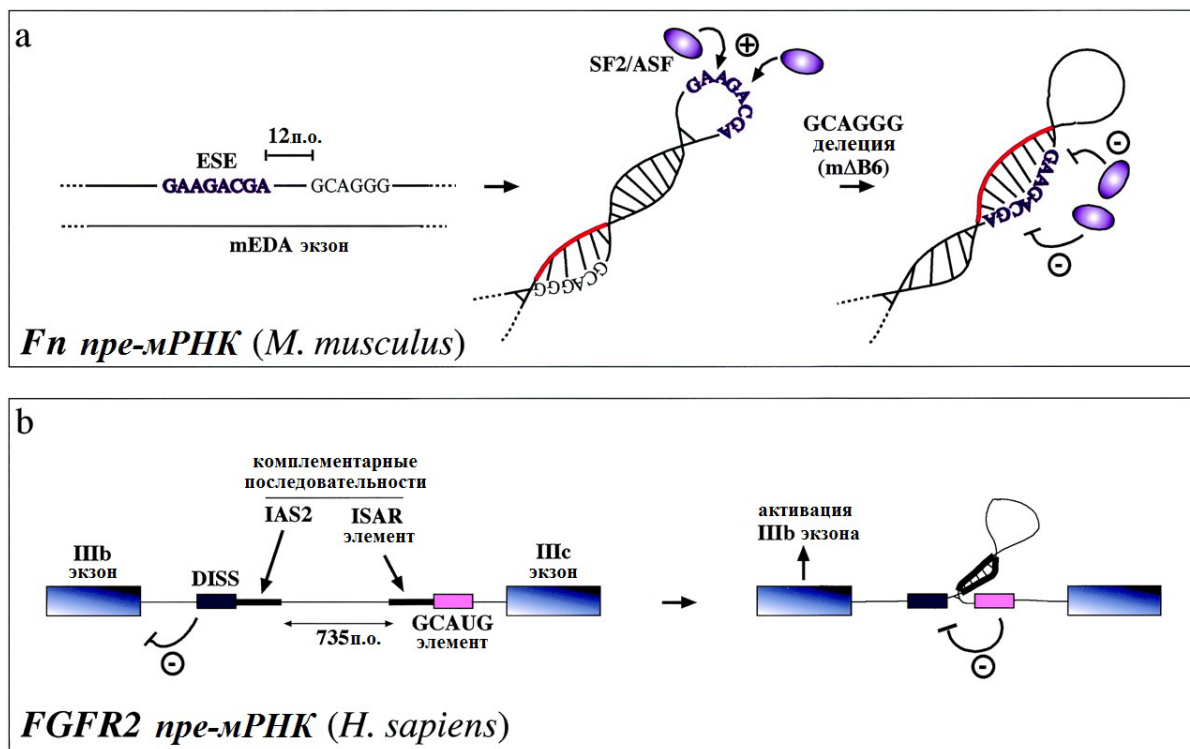


Рис. 10. Стерические изменения, оказывающие влияние на энхансеры и сайленсеры. (а) В экзоне фибронектина *EDA* человека сделана мутация, которая не нарушает последовательность экзонного энхансера, но перемещает его из петли в область комплементарных взаимодействий, что препятствует связыванию SF2/ASF и распознаванию экзона. (б) Элементы IAS2 и ISAR образуют вторичную структуру. Ее функция – приблизить контролирующий GCAUG-богатый элемент к интронному сайленсеру DISS, который подавляет включение экзона IIIb. Контролирующий элемент инактивирует DISS, что приводит к включению экзона IIIb в мРНК (по [3]).

ваторная последовательность, или Intronic Activating Sequence 2) и ISAR (интронный активатор и репрессор сплайсинга, или Intronic Splicing Activator and Repressor) с петлей длиной 735 нуклеотидов. Эта вторичная структура регулирует сплайсинг взаимоисключающих экзонов IIIb и IIIc, сближая контролирующий GCAUG-богатый элемент с интронным сайленсером DISS, который подавляет включение экзона IIIb. В результате происходит инактивация интронного сайленсера DISS и включение экзона IIIb (Рисунок 10b) [18]. Если бы вторичная структура не образовывалась, контролирующий элемент располагался бы слишком далеко от интронного сайленсера DISS, чтобы оказывать на него ка-

кое-нибудь влияние.

3. Регуляция сплайсинга с помощью белковых факторов

Как уже упоминалось выше, экзонные и интронные энхансеры и сайленсеры функционируют посредством связывания белковых транс-факторов сплайсинга, способных активировать или подавлять распознавание сайтов сплайсинга или сборку сплайсосомы.

Многие экзонные энхансеры содержат в себе сайты связывания регуляторных белков из семейства SR, которые принимают участие в нескольких этапах сборки сплайсосомы и могут функционировать одновременно в качестве факторов, необходимых для сплайсинга, и в качестве регуляторных факторов (Рисунок 11). Структура белков семейства SR содержит один или два N-концевых домена с РНК-распознающими мотивами (RRM-домены) и С-концевые домены, обогащенные остатками аргинина и серина (RS-домены), которые отвечают за белок-белковые и белок-РНКовые контакты [19]. RS-домены также присутствуют в других факторах сплайсинга, например в U2AF65 и U2AF35, а также в регуляторах альтернативного сплайсинга, таких как белок TRA у *D. melanogaster*.

Экзонные и интронные сайленсеры сплайсинга имеют переменные последовательности, однако по крайней мере некоторые из них связывают белки из обширного семейства hnRNP. Белки семейства hnRNP содержат RRM домены и KH-домены (домены, которые связывают одноцепочечные РНК), а также вспомогательные домены, которые отвечают за белок-белковые взаимодействия. Белки семейства hnRNP являются одними из наиболее высоко экспрессированных в клетке белков и выполняют самые разнообразные функции в метаболизме пре-мРНК: упаковка только что синтезированных транскриптов; регуляция конститутивного и альтернативного сплайсинга; транспорт молекул

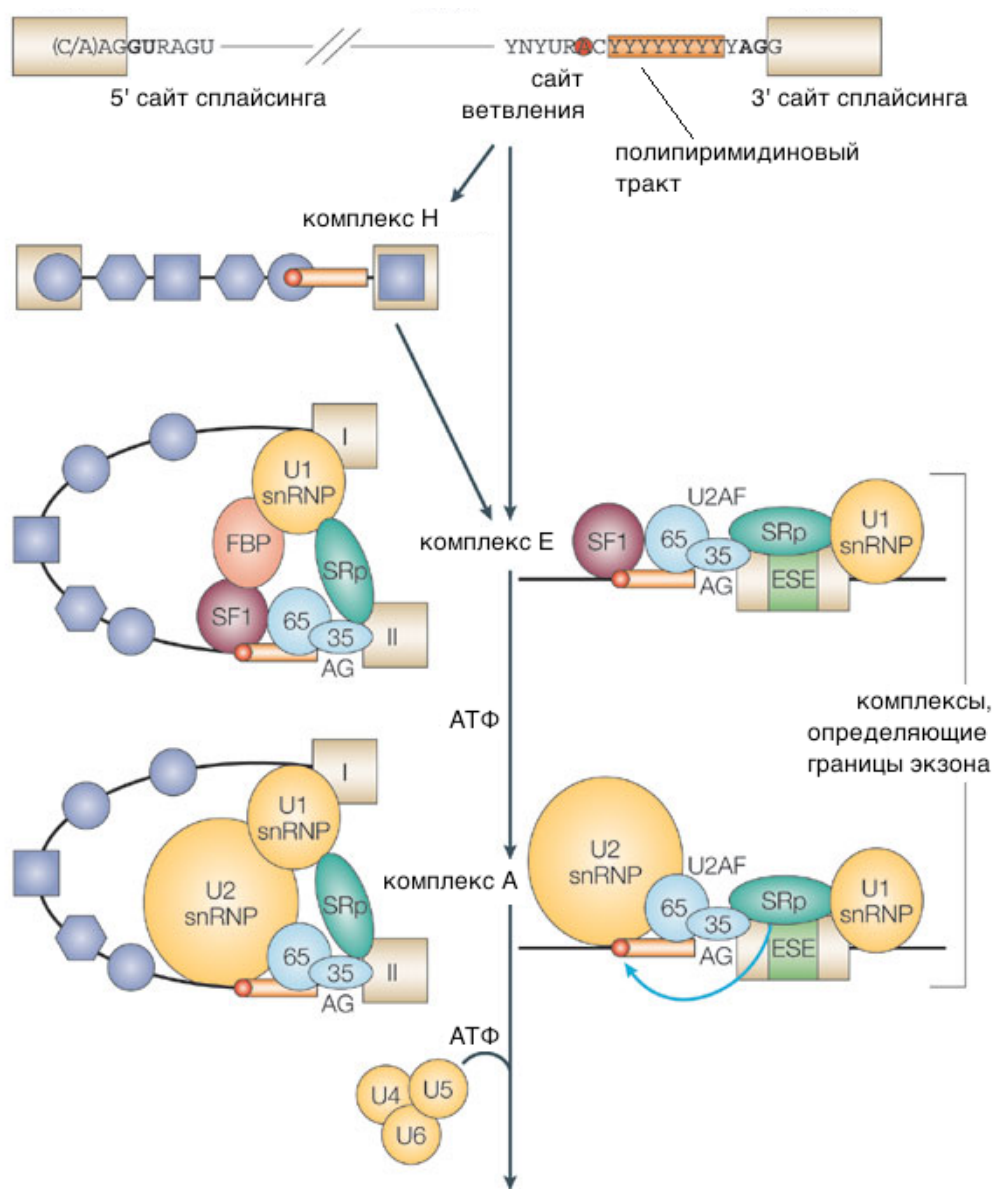


Рис. 11. Сборка сплайсосомы на пре-мРНК проходит в несколько этапов. (1) Комплекс Н с белками семейства hnRNP (обозначены как фигуры различной формы голубого цвета). (2) Комплекс Е, в котором все консенсусные сигналы сплайсинга связаны с компонентами сплайсосомы, а экзонные энхансеры - с белками семейства SR. (3) Комплекс А, в котором происходит связывание мРНК U2. (4) Комплекс В, образующийся в результате сближения сайтов сплайсинга и присоединения оставшихся компонентов сплайсосомы (по [20]).

мРНК и их локальная трансляция; регуляция стабильности мРНК; активация или репрессия трансляции [21].

Некоторые из белков семейства hnRNP выступают в качестве репрессоров

сплайсинга (белки человека hnRNPA1 и PTB/hnRNPI; белок SXL *D. melanogaster*). Репрессоры, как и активаторы сплайсинга, обычно функционируют посредством влияния на образование комплексов E и A на начальных этапах сборки сплайсосомы (Рисунок 11). Комплекс H также выполняет регуляторную функцию за счет связывания белков из семейства hnRNP, которые могут способствовать или препятствовать взаимодействию компонентов сплайсосомы с альтернативными сайтами сплайсинга.

3.1. Механизмы активации сплайсинга

Механизм активации 3'- и 5'-сайтов сплайсинга в результате связывания белков SR с экзонными энхансерами изучен довольно хорошо. Упрощенная модель этого механизма, опосредованного присоединением белка U2AF65 к слабому полипиримидиновому тракту, представлена на рис. 12а. Примером такой

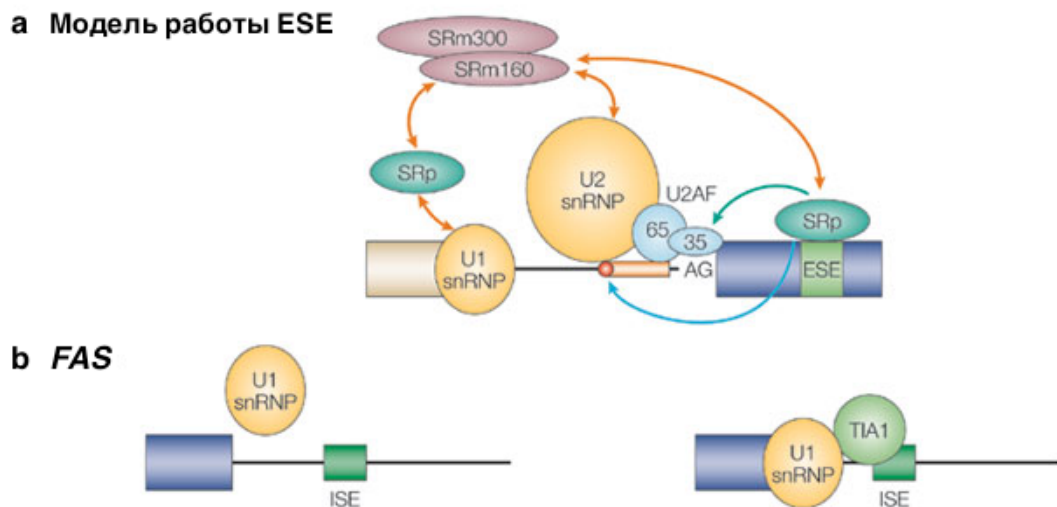


Рис. 12. (а) Модель работы экзонного энхансера (ESE). SR-белки (SRp), связываясь с экзонным энхансером, могут активизировать сплайсинг посредством притягивания белка U2AF65 к слабым полипиримидиновым трактам в результате связывания их RS-доменов с белком U2AF35 (зеленая стрелка), взаимодействия с ко-активатором SRm160 (оранжевые стрелки) или контакта с сайтом ветвления (голубая стрелка). (б) Действие слабого 5'-сайта сплайсинга в транскрипте FAS усиливается в результате связывания белка TIA1 с интронным энхансером (ISE). TIA1 усиливает взаимодействие мРНК U1 с пре-мРНК (по [20]).

активности является пол-специфичный экзонный энхансер в экзоне 4 гена *dsx* у *D. melanogaster*, который состоит из последовательности, богатой пуринами, и шести 13-нуклеотидных повторов (мотивов dsxRE). Каждый повтор связывает белок SR вместе с двумя регуляторами, содержащими RS-домены, – TRA и TRA2. TRA выступает в качестве пол-специфичного регулятора, способствуя стабилизации энхансерного комплекса. Этот комплекс действует посредством взаимодействия RS-доменов со вспомогательным белком U2AF35, таким образом способствуя связыванию U2AF65 со слабым полипиримидиновым трактом [22].

Существует альтернативный механизм действия энхансеров, опосредованный ко-активаторами сплайсинга SRm160 и SRm300 (Рисунок 12а). Эти ко-активаторы также содержат RS-домены, но RRM-домены у них отсутствуют. Они могут взаимодействовать с различными мРНК и с белками SR, связанными с энхансерами. Экзонные энхансеры также могут активировать 5'-сайты сплайсинга. Пол-специфичный выбор альтернативного 5'-сайта сплайсинга в гене *fruitless* у *D. melanogaster* зависит от энхансера в предшествующем экзоне, который содержит три копии 13-нуклеотидного повтора dsxRE, который связывает белки TRA и TRA2. Эксперименты показали, что в данном случае TRA и TRA2 действуют так же, как и для 5'-сайта сплайсинга (см. выше) [23].

Рассмотренные модели работы экзонных энхансеров основаны на способности RS-доменов обеспечивать белок-белковые взаимодействия. Недавние исследования показали, что с сайтом ветвления и 5'-сайтом сплайсинга может последовательно связываться целая серия RS-доменов [24]. В частности, RS-домены, которые связываются с экзонными энхансерами, также могут контактировать с сайтом ветвления в комплексе A. Сильная корреляция между способностью активировать сплайсинг и контактировать с сайтом ветвления, наблюдаемая в комплексе A, говорит о том, что взаимодействие между RS-доменами и сайтом ветвления может лежать в основе механизма активации сплайсинга экзонными

энхансерами.

Белки семейства hnRNP, которые обычно считаются репрессорами, могут также стимулировать использование альтернативных сайтов сплайсинга, главным образом, через интронные энхансеры. Так, активация слабого 5'-сайта сплайсинга в гене *eNOS* регулируется связыванием белка hnRNPL с интронным энхансером, состоящим из CA-повторов [25]. Близкородственные белки TIA1 и TIAR связываются с урацил-богатыми интронными энхансерами, чтобы способствовать взаимодействию мРНК U1 со слабым 5'-сайтом сплайсинга, видимо за счет прямого контакта с мРНК U1-специфичным C-белком (Рисунок 12b).

3.2. Механизмы репрессии сплайсинга

Репрессия сплайсинга часто осуществляется белками семейства hnRNP, такими как SXL, PTB и hnRNPA1. В самом простом случае, белки, связываясь с сайленсерами, напрямую препятствуют связыванию факторов сплайсинга. Однако, для репрессии часто требуется кооперативное связывание нескольких репрессоров с несколькими сайленсерами одного типа. Две модели кооперативной репрессии предполагают создание так называемой зоны сайленсинга, либо за счет расширения зоны связывания репрессоров от высоко-аффинных сайленсеров, либо за счет выпетливания РНК между связанными репрессорами. В любом случае, экзоны, сайты сплайсинга или энхансеры становятся недоступными для факторов сплайсинга.

Примером простого стерического ингибирования сборки сплайсосомы на раннем этапе является ген *tra* у *D. melanogaster*. У самок белок *SXL* связывается с полипиримидиновым трактом, блокируя таким образом доступ фактора U2AF. В результате выбирается альтернативный 3'-сайт сплайсинга, с которым связывается белок U2AF, а не SXL (Рисунок 13a). Подавление сплайсинга белком PTB также происходит за счет конкуренции с U2AF65. Однако, в данном

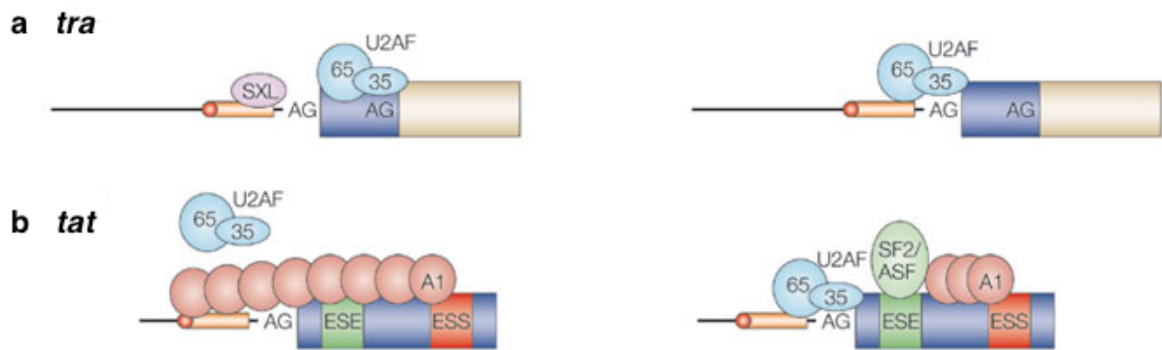


Рис. 13. (а) Репрессия 3'-сайта сплайсинга в гене *tra* в результате взаимодействия белка SXL с интронным сайленсером (ISS), расположенным внутри полипиримидинового тракта, которое препятствует связыванию белка U2AF и приводит к использованию альтернативного 3'-сайта сплайсинга. (б) Уровень включения экзона 3 в гене HIV1 *tat* определяется соотношением белков из семейств hnRNP и SR в ядре. Взаимодействие мультимеров белка hnRNPA1 с экзонным сайленсером (ESS) стерически блокируется в результате взаимодействия белка SF2/ASF с экзонным энхансером (ESE). В данном случае, функционирование экзонного энхансера требует связывания RRM-домена, а не RS-домена белка SF2/ASF (по [20]).

случае, как правило, требуется кооперативное связывание РТВ с несколькими сайтами, часто не имеющими отношения к полипиримидиновому тракту. Авторегуляция сплайсинга гена *Sxl* происходит за счет кооперативного связывания белка SXL к нескольким сайтам вокруг экзона 3, что ведет к пропуску этого экзона [20].

Белок hnRNPA1 функционирует как общий регулятор альтернативного сплайсинга, не зависящий от последовательности и являющийся антагонистом белков SR, таких как SF2/ASF74. Вариации соотношения SF2/ASF и hnRNPA1, происходящие в результате изменения внутриклеточной локализации, могут оказывать влияние на многие события альтернативного сплайсинга (Рисунок 13б). Фосфорилирование белка hnRNPA1 в результате стресса и запуска МКК3/6–p38 сигнального каскада приводит к экспорту hnRNPA1 в цитоплазму, с последующими изменениями альтернативного сплайсинга в ядре. Высоко-аффинные сайты связывания hnRNPA1 функционируют как специфические сайленсеры.

Механизм подавления сплайсинга зависит от локализации сайтов связывания и может происходить за счет конкуренции с мяРНП U1, U2 или белками SR.

В случае авторегуляции гена *hnRNPA1*, пропуск экзона 7В происходит в результате взаимодействия высоко-аффинных сайтов связывания белка hnRNPA1, расположенных по обе стороны от экзона 7В (Рисунок 14). Интересно, что данное взаимодействие обеспечивается как непосредственным связыванием двух белков hnRNPA1 друг с другом, так и формированием вторичной структуры РНК между их сайтами связывания. В результате происходит выпетливание экзона 7В, сайты сплайсинга внешних экзонов сближаются и получают стерическое преимущество [3].

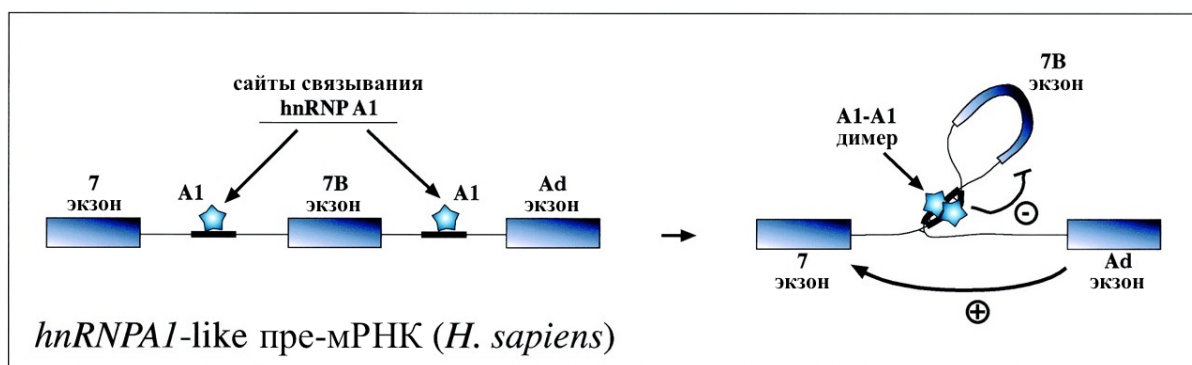


Рис. 14. Выпетливание и ингибирование включения экзона в гене *hnRNPA1* происходит за счет белков hnRNPA1, которые связываются по обе стороны от экзона и взаимодействуют друг с другом. Кроме того, их сайты связывания образуют между собой вторичную структуру РНК. Сайты сплайсинга внешних экзонов сближаются и получают преимущество (по [3]).

3.3. Регуляция спрайсинга посредством белка hnRNPL

Белок hnRNPL – это ещё один транс-фактор регуляции сплайсинга из семейства белков hnRNP. Функция его до конца не ясна, однако известно, что белок hnRNPL связывается преимущественно с СА-повторами. Также существуют экспериментальные данные, свидетельствующие о значительном влиянии hnRNPL на процесс сплайсинга в целом. Результаты нескольких исследований указывают на то, что hnRNPL регулирует пропуск экзона в гене *CD45*, в ответ

на активацию Т-клеток, возможно через стабилизацию связывания мРНК [26, 27]. Белок hnRNPL также регулирует альтернативный сплайсинг других генов посредством удерживания интрона, подавления включения нескольких экзонов и выбор альтернативного сайта полиаденилирования. Белок hnRNPL взаимодействует с 3'UTR мРНК синтазы оксида азота и регулирует её стабильность [21].

Сайты связывания белка hnRNPL могут быть определены экспериментально с помощью метода, называемого CLIP. Этот метод позволяет идентифицировать белки, которые напрямую контактируют с РНК, а также места контакта в живых клетках на полногеномном уровне. Метод CLIP включает в себя следующие этапы [28]:

1. Облучение клеток или тканей ультрафиолетом, чтобы образовать ковалентные связи между РНК и связанным с ней белком.
2. Лизис клеток и частичное расщепление РНКазой.
3. Иммунопреципитация изучаемого белка, вместе со связанной с ним РНК.
4. Мечение 5'-концов фрагментов РНК и очищение РНК-белковых комплексов электрофорезом.
5. Удаление белков с помощью протеазы и выделение РНК фрагментов.
6. Лигирование адапторов и полимеразная цепная реакция после обратной транскрипции РНК.
7. Секвенирование.

Технология CLIP позволяет очистить РНК-белковые комплексы с высокой точностью, без разрушения комплексов, и идентифицировать участки пре-мРНК, с которыми связывается белок, с точностью до одного нуклеотида.

4. Транс-сплайсинг

Механизмы сплайсинга пре-мРНК, управляемые мажорной и минорной формами сплайсосомы, – не единственные известные к настоящему времени формы процессинга РНК эукариот, приводящие к объединению и лигированию экзонов. Существует и третий механизм сплайсинга – транс-сплайсинг, особая форма процессинга РНК, в результате которой экзоны, находящиеся на двух разных молекулах РНК, стыкуются и лигируются. Транс-сплайсинг был впервые обнаружен в 1986 г. у трипаносом, затем у *Caenorhabditis elegans* и других нематод, простейших *Euglena*, плоских червей и даже у высших эукариот [29].

В геномах нематод и трипаносом присутствуют короткие последовательности, называемые 'лидерами сплайсинга' ('Splice Leader', или SL-последовательностями). В результате транскрипции и последующего транс-сплайсинга SL-последовательности присоединяются к 5'-концам многих белок-кодирующих генов (Рисунок 15). SL-последовательности происходят из специальных генов, называемых SL-генами, которые могут располагаться за тысячи нуклеотидов, или даже на других хромосомах по отношению к тем белок-кодирующим генам, к 5'-концам которых они присоединяются. Интересно, что при транс-сплайсинге сайты сплайсинга и сайт ветвления не обладают такой выраженной консервативностью, как при цис-сплайсинге. При этом возрастает значимость полипиримидинового тракта. У трипаносомы он существенно длиннее (около 20 нуклеотидов), чем у млекопитающих (10-12 нуклеотидов), и начинается сразу за сайтом ветвления [30].

Транс-сплайсинг встречается не только у трипаносом и червей. Имеются сведения и о транс-сплайсинге в клетках человека. Так, транс-сплайсинг может происходить между 5'-экзонами гена JAZF1 в локусе 7p15 и 3'-экзонами гена JJAZ1 (также известного как SUZ12) в локусе 17q11, причем получившийся в результате химерный транскрипт транслируется в белок, препятствующий

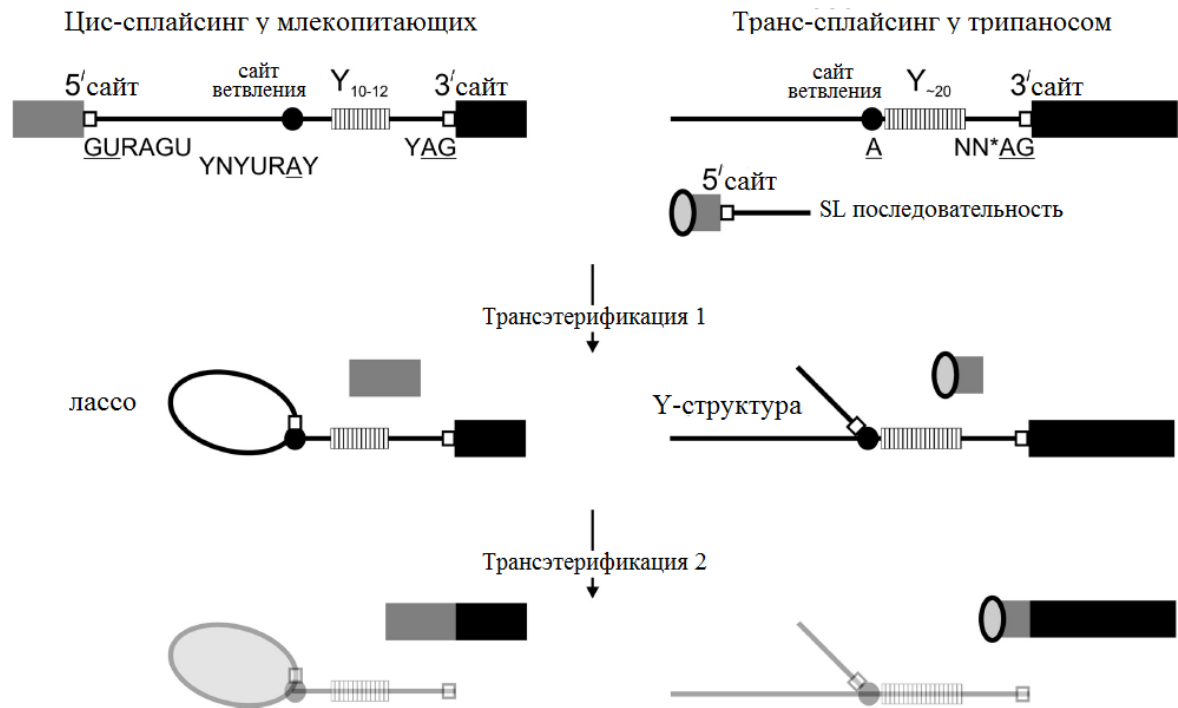


Рис. 15. Сравнение цис-сплайсинга млекопитающих и транс-сплайсинга нематод. Первый экзон и SL-последовательность изображены серыми прямоугольниками; второй экзон и белок-кодирующий ген трипаномы – черными прямоугольниками; сайты сплайсинга – незакрашенными квадратами; сайт ветвления – черным кругом. После первой реакции трансэтерификации, при цис-сплайсинге интрон образует структуру в форме петли, а при транс-сплайсинге образуется структура другой формы. После второй реакции трансэтерификации, обе этих интронных структуры разворачиваются и быстро деградируют (по [30]).

апоптозу [31]. Транс-сплайсинг наблюдается и между генами SLC45A3 и ELK4, также с образованием функционального белка [32].

4.1. Химерные РНК

Секвенирование транскриптомов различных организмов, от червей до человека, показало, что многие транскрипты состоят из сегментов последовательностей, которые не следуют друг за другом на хромосоме в базовой версии генома, а происходят из удаленных участков генома, и даже с разных хромосом. Некоторые из таких химерных транскриптов являются артефактами экспериментальной процедуры секвенирования [33], а остальные образуются в

результате генетических перестроек у объекта секвенирования (чаще всего, в опухолевых тканях) [34], либо возникают в нормальных тканях пост-транскрипционно в результате транс-сплайсинга [29]. Это наблюдение указывает на то, что транс-сплайсинг у млекопитающих - не редкое явление, как считалось ранее, а довольно распространенный механизм.

Обычно химерные транскрипты детектируют с помощью парно-концевого секвенирования транскриптома, при котором молекулы РНК превращают в кДНК, разрезают на фрагменты длиной в несколько сотен нуклеотидов, а затем определяют, 'считывают' последовательности фрагментов с обоих концов. Полученные парные прочтения имеют, как правило, строго определенную длину. Их выравнивают на геном и отбирают такие пары, прочтения которых выравниваются на разные гены (Рисунок 16). Эти транс-пары прочтений происходят из химерных транскриптов, имеющих в своем составе участки последовательности, принадлежащие двум разным генам. Транс-пары обычно составляют 1-5% от всех пар прочтений, полученных в результате секвенирования образцов как раковых, так и нормальных тканей, что позволяет говорить о потенциальной функциональной значимости химерных транскриптов, а значит и транс-сплайсинга.

4.2. Систематические ошибки секвенирования

Существует мнение, что наблюдаемые химерные транскрипты могут быть следствием какого-либо неизученного артефакта процедуры секвенирования [33], хотя в той же работе с помощью независимого от секвенирования эксперимента было идентифицировано 80 генов, в которых у *Drosophila* происходит транс-сплайсинг между гомологичными аллелями.

Много усилий было приложено, чтобы обнаружить, понять причину и устранить систематические ошибки, присущие современным технологиям высо-

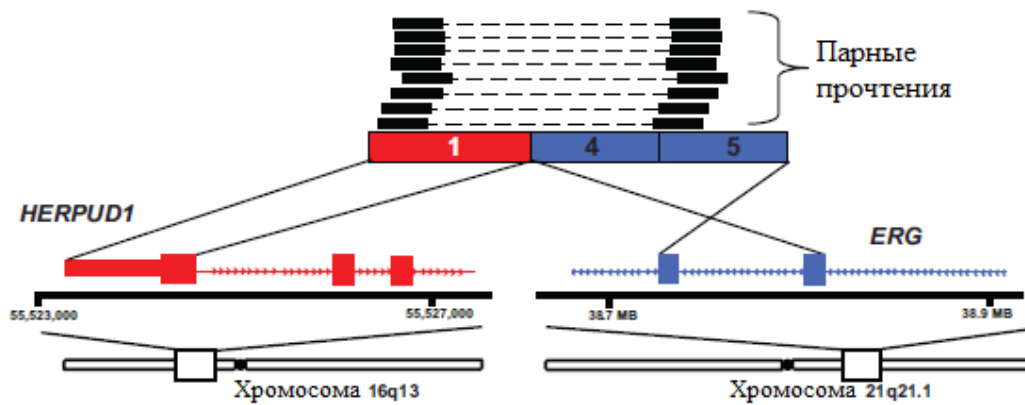


Рис. 16. Поиск ранее неизвестных химерных транскриптов в ткани рака простаты. Схематически изображен химерный транскрипт между экзоном 1 гена *HERPUD1* (красный), расположенного на хромосоме 16, и экзоном 4 гена *ERG* (синий), расположенного на хромосоме 21 (по [34]).

копроизводительного секвенирования. Хотя и не все систематические ошибки удалось эффективно скорректировать, найдено их было много. Так, для платформ компании Иллюмина (Illumina) характерно следующее.

1. Участки последовательности, имеющие повышенное содержание гуанина и цитозина, покрыты большим количеством прочтений, чем последовательности с нормальным нуклеотидным составом.
2. Ошибки секвенирования встречаются преимущественно вблизи 3'-концов прочтений.
3. Последовательности, предшествующие ошибкам секвенирования, обогащены гуанином.
4. Трансверсии $G \rightarrow T$ и $A \rightarrow C$ встречаются чаще, чем другие ошибки.
5. Оценки качества секвенирования завышены для позиций с высоким качеством и занижены для позиций с низким качеством [35].
6. Прочтения распределены неравномерно вдоль секвенируемых транскрип-

тов из-за отличий в частоте встречаемости нуклеотидов на разных участках этих транскриптов [36, 37].

7. Идентичные фрагменты кДНК перепредставлены из-за повышенной амплификации в ходе полимеразной цепной реакции (ПЦР) [38].
8. Участки генома с низкой сложностью последовательности хуже покрыты прочтениями из-за трудностей с выравниванием прочтений.
9. 5'-концы транскриптов имеют повышенное покрытие из-за плохого гидролиза [39].
10. Примесь непротранскрибированных или недосплайсированных транскриптов приводит к повышению покрытия интронов.
11. Дегградация мРНК в ходе отбора полиаденилированных транскриптов ведет к постепенному повышению покрытия в сторону 3'-конца транскрипта.
12. Вторичная структура РНК может оказывать влияние на фрагментацию.

Учитывая, что систематические ошибки столь распространены и разнообразны, остается открытым вопрос, можно ли доверять результатам секвенирования, и если да, то как использовать данные секвенирования так, чтобы негативное влияние систематических ошибок было минимальным.

4.3. Трехмерная архитектура генома человека

Недавние исследования показали, что удаленные участки ДНК часто контактируют между собой, даже если они располагаются на разных хромосомах [40]. И эти контакты не случайны, а, напротив, статистически значимы и, по видимому, функциональны. Согласно современным представлениям, хромосо-

мы в интерфазном ядре организованы в достаточно четко выраженные хромосомные территории [41]. Богатые генами хромосомы занимают более внутренние позиции в ядре, в то время как бедные генами хромосомы располагаются на периферии. При этом активно транскрибирующиеся гены чаще располагаются на поверхности хромосомных территорий. Ко-регулируемые гены могут образовывать внутри- и межхромосомные контакты, формируя так называемые транскрипционные фабрики, обогащенные транскрипционными факторами и факторами сплайсинга. Внутри- и межхромосомные контакты участвуют в регуляции экспрессии генов, доставляя регуляторные белки к регулируемым ими генам [42].

Поскольку механизм транс-сплайсинга предполагает физическое объединение двух транскриптов и предварительное связывание факторов сплайсинга с каждым из них, то логично предположить, что пространственно близкие участки разных хромосом, принадлежащие одной транскрипционной фабрике, должны образовывать большее количество химерных транскриптов, чем пространственно удаленные (Рисунок 17).

Развитие технологии определения конформации хромосом (3C) позволило детально изучить дальние взаимодействия в хроматине. Этот метод основан на пространственно ограниченном лигировании, совмещенном с локус-специфичной полимеразной цепной реакцией [43]. Недавно была разработана новая технология, которая называется Hi-C. Она позволила расшифровать трехмерную архитектуру всего генома человека посредством лигирования пространственно близких участков ДНК и последующего секвенирования лигированных фрагментов [40]. Была построена полногеномная матрица пространственной близости фрагментов генома человека с разрешением в 1 миллион нуклеотидов (1Mb).

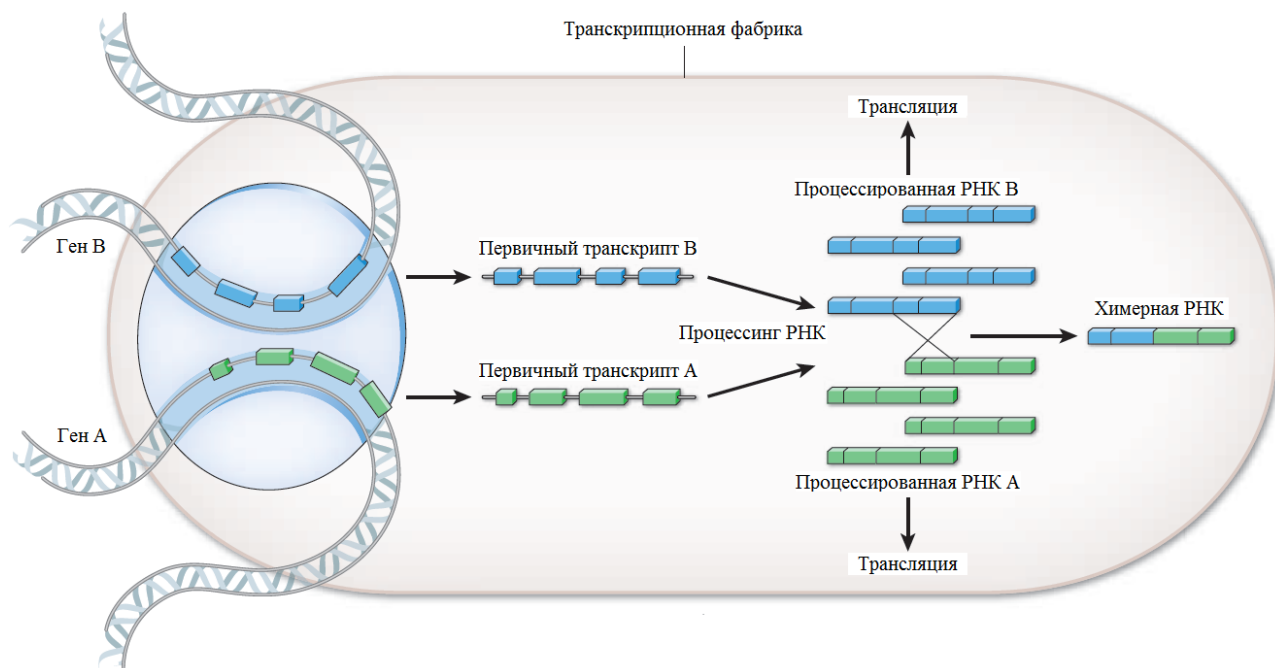


Рис. 17. Модель транскрипционной фабрики, синтезирующей химерные РНК. Предполагается, что гены А и В, расположенные на разных хромосомах, объединены в транскрипционную фабрику. Пре-мРНК, синтезируемые с этих генов, процессируются в зрелые мРНК и химерные РНК. При этом большая часть пре-мРНК подвергается цис-сплайсингу, и лишь небольшая доля пре-мРНК превращается в химерные транскрипты. Поскольку гены А и В пространственно сближены, между ними могут чаще происходить генетические перестройки, которые также являются источником химерных РНК (по [29]).

4.4. Эпигенетические свойства генома человека

Поскольку пространственно сближенные гены, принадлежащие одной транскрипционной фабрике, по всей видимости, являются ко-регулируемыми и пользуются одними и теми же транскрипционными факторами и факторами сплайсинга, то представляется интересным выяснить, насколько похожи функции данных генов.

Современные технологические достижения позволили получить огромное количество данных о функциональном состоянии хроматина посредством эпигенетических маркеров, таких как метилирование ДНК, модификации гистонов, чувствительность к ДНКазе (отражает открытое состояние хроматина) [44–46].

Было показано, что эти характеристики коррелируют друг с другом, а также с локализацией генов, промоторов, энхансеров и других функциональных участков, а также с уровнем экспрессии генов [47–50]. Также доступна полногеномная функциональная аннотация генома человека, содержащая координаты различных классов комбинаций эпигенетических маркеров [51]. Такие классы комбинаций эпигенетических маркеров называют состояниями хроматина.

5. Выводы

1. Альтернативный сплайсинг — это один из основных механизмов создания разнообразия белковых последовательностей, подверженный сложной регуляции.
2. Экспериментальные подтверждения влияния вторичной структуры РНК на связывание факторов сплайсинга говорят о значительном участии вторичной структуры в процессе регуляции сплайсинга.
3. Пластичность сплайсинга обеспечивается наличием регуляторных сигналов - энхансеров и сайленсеров, которые функционируют посредством связывания белковых факторов.
4. Известно несколько примеров регуляции сплайсинга белком hnRNPL, одним из высоко экспрессированных в клетке белков. Влияние hnRNPL на сплайсинг не изучено на полногеномном уровне.
5. Транс-сплайсинг - это особая форма процессинга РНК, в результате которой экзоны, находящиеся на двух разных молекулах РНК, стыкуются и лигируются.
6. Современные данные секвенирования указывают на то, что транс-сплайсинг у млекопитающих — нередкое явление, возможно связанное с про-

странственной структурой хроматина.

Глава 1

Поиск вторичных структур РНК, участвующих в регуляции сплайсинга

1.1. Материалы и методы

1.1.1. Объекты исследования

В настоящей работе поиск консервативных вторичных структур был выполнен для группы плацентарных млекопитающих: *H. sapiens*, *R. macaque*, *P. troglodytes*, *M. musculus*, *R. norvegicus*, *C. familiaris*, *F. catus*, *B. taurus*, *E. caballus*, *M. domestica*. Для них была составлена база данных событий сплайсинга на основе данных RefSeq, полученных из UCSC genome browser database. Она содержала около 383 тыс. сайтов сплайсинга, 200 тыс. интронов и 213 тыс. экзонов. Ортологи сайтов сплайсинга человека в других видах были также получены из UCSC genome browser database.

1.1.2. Подходы к решению задачи

В настоящей работе была поставлена задача поиска элементов вторичных структур вблизи сайтов сплайсинга. Каковы возможные пути решения этой задачи? Можно осуществить поиск комплементарных участков по геному одного организма. Однако применение такой процедуры даст слишком большое количество предполагаемых вторичных структур, и среди них будут представлены как имеющие биологический смысл структуры, так и комплементарные по случайным причинам последовательности. Поэтому применяют сравнительно-геномный подход, который предполагает поиск консервативных вторичных структур по последовательностям нескольких эволюционно родственных организмов. И

если какая-либо структура присутствует в нескольких организмах (т.е. консервативна), то можно надеяться, что такая находка не случайна и имеет биологический смысл. Поэтому в настоящей работе нас интересовали только одновременно комплементарные и консервативные последовательности.

Такие структуры можно найти двумя способами (Рисунок 1.1): (1) определить комплементарные участки и выбрать из них только консервативные; (2) построить выравнивание, затем найти в нем комплементарные участки. В настоящей работе реализован первый подход, в основе которого лежит поиск достаточно протяженных комплементарных последовательностей нуклеотидов ('затравок'). Далее, среди затравок отбираются эволюционно консервативные последовательности, которые расширяются до более протяженных комплементарных участков [52]. Технически данный метод был реализован с помощью хэширования.

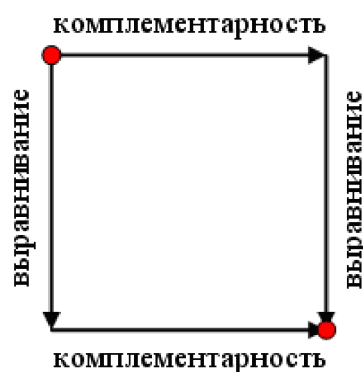


Рис. 1.1. Методы решения задачи поиска консервативных вторичных структур. Проверять последовательности на консервативность и комплементарность можно в разном порядке, поэтому возникает два пути решения задачи, как показано на схеме.

1.1.3. Поиск консервативных вторичных структур с помощью хэширования

Хэширование – это подход, широко применяемый в различных алгоритмах сравнения биологических последовательностей. Он используется во мно-

гих биоинформатических программах, таких как, например, *BLAST*. В основе хэширования лежит построение словарей. Для каждого генома, участвующего в процедуре поиска консервативных вторичных структур, составляются два словаря: для окружений донорных и акцепторных сайтов сплайсинга. Словарь представляет собой хэш-таблицу, в которой ключами являются все возможные слова заданной длины, и каждому слову сопоставлен список интронов, в которых данное слово встречается. В словаре для донорных частей интронов слова упорядочены по алфавиту, а в словаре для акцепторных частей интронов слова отсортированы в комплементарном порядке. В результате, в соответствующих по номеру позициях донорного и акцепторного словаря оказываются комплементарные слова (Таблица 1.1).

Далее для составленных словарей находили теоретико-множественное пересечение: последовательно сравнивали списки интронов, соответствующих комплементарным словам (словам, стоящим под одинаковыми номерами в словарях). Если интрону из первого списка не находилось пары во втором списке, то он удалялся из словаря. Интроны из второго списка, для которых не нашлось пары в первом списке, также удаляли. Эту процедуру пересечения производили отдельно для каждого организма.

Если при поиске комплементарных слов допускаются GT пары, одному слову в словаре для донорных частей интронов может соответствовать несколько слов в словаре для акцепторных частей интронов, и наоборот. То есть под одним номером в словарях может стоять несколько слов. Список интронов, в которых встречаются эти слова, расширяется соответственно (Таблица 1.2). Процедура пересечения словарей осуществляется точно так же, как и в случае поиска консервативных вторичных структур без учета GT пар.

После процедуры пересечения словарей осуществляли дополнительную проверку на консервативность, а именно на встречаемость пар интронов с комплементарными словами в геномах нескольких организмов. Сравнивали словари

Таблица 1.1. Пример словарей, используемых для поиска консервативных вторичных структур без учета GT пар. Словари составлены из всех возможных слов длины 2 и упорядочены таким образом, что в соответствующих по номеру позициях словарей находятся комплементарные слова. Каждому слову сопоставлен список интронов, в которых это слово встречается. В последней колонке представлен результат процедуры пересечения словарей (см. описание в тексте).

#	Донорный словарь		Акцепторный словарь		Пересечение
1	aa	<u>I0223</u> , I0015, I0665	tt	<u>I0223</u> , I0067	<u>I0223</u>
2	ac	I0543, I1005	gt	I0385, I0456, I0546	
3	ag	I0422, <u>I0654</u>	ct	I0444, <u>I0654</u>	
4	at	I0011, I0542, I0126	at	I0657, I0758	
5	ca	I0874	tg		<u>I0501</u>
6	cc	I0091, <u>I0501</u> , I1042	gg	I0739, <u>I0501</u> , I1009	
7	cg		cg	I0937, I0013	
8	ct	I0757, I0672	ag	I0700, I0723	
9	ga	I0854, I0748, I0639	tc	I0659	<u>I0843</u>
10	gc	<u>I0843</u>	gc	<u>I0843</u> , I0396, I0593	
11	gg	I0753, I0760	cc	I0493	
12	gt	<u>I0530</u> , I0739, <u>I0367</u>	ac	<u>I0530</u> , I1015, <u>I0367</u>	
13	ta		ta	I0932, I0111, I0217	<u>I0530</u> , <u>I0367</u>
14	tc	I0473	ga	I0513	
15	tg	I0483, I0753, I0743	ca		
16	tt	I0675, <u>I0845</u>	aa	I0014, <u>I0845</u>	

для разных организмов и удаляли из них те пары интронов, которые присутствовали в меньшем числе геномов, чем задано. В результате такой процедуры получали словарь, содержащий искомые вторичные структуры, удовлетворяющие требованиям комплементарности и консервативности одновременно. Необходимо заметить, что выравнивание в данном подходе не используется, что позволяет избежать ошибок, связанных с неправильным построением выравнивания. Выравнивание применяется только для визуализации найденных структур.

Таблица 1.2. Пример словарей, используемых для поиска консервативных вторичных структур с учетом GT пар. Все обозначения как в Таблице 1.1.

#	Донорный словарь		Акцепторный словарь		Пересечение
1	aa	I0223, I0015, I0665	tt	I0223, I0067	I0223
	ga	I0854, I0748, I0639			
	ag	I0422, I0654			
2	ac	I0543, I1005	gt	I0385, I0456, I0546	
	at	I0011, I0542, I0126			
	gc	I0843			
...	...	...	...	...	...
15	tg	I0483, I0753, I0743	ca		
			ta	I0932, I0111, I0217	
			cg	I0937, I0013	
16	tt	I0675, I0845	aa	I0014, I0845	I0845
			ga	I0513	
			ag	I0700, I0723	

1.1.4. Контроль процедуры поиска вторичных структур

Классическим методом определения статистической значимости предсказанных вторичных структур РНК является сравнение полученных структур со структурами, полученными в результате применения того же алгоритма к набору последовательностей с перемешанными нуклеотидами [53]. В нашем случае использование этого метода вызывает трудности, поскольку он требует согласованного перемешивания нуклеотидов во всех организмах, которое невозможно осуществить без множественного выравнивания последовательностей. Кроме того, такое перемешивание влияет на энергию вторичных структур, которая зависит от динуклеотидного состава последовательности. Поэтому уровень ложных положительных предсказаний был оценен при помощи другого метода, в ходе которого перемешивались акцепторные части интронов (одновременно во всех видах для ортологичных сайтов сплайсинга). Перемешивание осуществлялось

таким образом, чтобы каждая новая пара донор-акцептор была составлена из интронов, принадлежащих разным генам. В результате каждой донорной части интрона оказалась сопоставленной акцепторная часть чужого интрона (Рисунок 1.2).

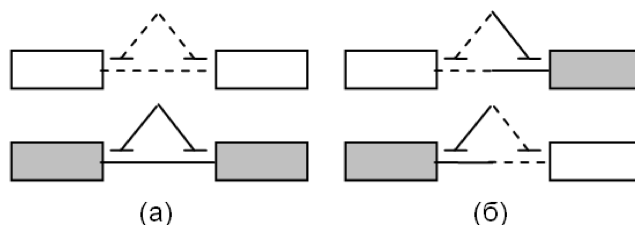


Рис. 1.2. Контроль процедуры поиска консервативных вторичных структур. (а) Области поиска консервативных вторичных структур в опыте. (б) Аналогичные области в контроле: каждой донорной половине интрона ставится в соответствие чужая акцепторная.

Далее для каждого искусственно составленного интрона была проведена обычная процедура поиска консервативных вторичных структур (см. выше). Процедура искусственного составления интронов затрагивает только хэш таблицы, не меняя исходные последовательности, что обеспечивает быстроту расчетов. Поэтому мы повторили контрольную процедуру 1000 раз, получив на каждом шаге число искусственно составленных интронов, в которых найдены консервативные вторичные структуры. По этим данным был вычислен уровень ложных положительных предсказаний как отношение числа вторичных структур, найденных в контроле, к числу вторичных структур, найденных в опыте.

Поскольку вероятность нахождения консервативной вторичной структуры зависит от нуклеотидного состава и консервативности последовательностей, мы ввели дополнительные ограничения при перемешивании, чтобы акцепторные части интронов обменивались только в том случае, если они имеют схожий GC-состав и/или уровень консервативности. Сайты сплайсинга были разделены на 10 групп по среднему GC-составу (от 0% до 100%), и перемешивание акцепторных частей интронов осуществлялось только внутри группы. Схожий уровень

консервативности обеспечивался таким же способом: акцепторные части интронов перемешивались только внутри группы с такой же консервативностью. Консервативность была оценена без использования множественных выравниваний по содержанию к-меров, как описано в [54]. Таким образом мы учли разные вероятности формирования вторичных структур в АТ-богатых и GC-богатых последовательностях, а также тот факт, что вероятность нахождения консервативной вторичной структуры выше в последовательностях с высокой средней консервативностью.

1.2. Результаты и обсуждение

Мы предполагаем, что вторичные структуры, расположенные вблизи сайтов сплайсинга, вероятнее ассоциированы со сплайсингом, чем удаленные структуры. Поэтому (а также для уменьшения вычислительной сложности) для поиска консервативных вторичных структур мы рассматривали только окружения донорных и акцепторных сайтов сплайсинга. При этом под окружением понимались 150 нуклеотидов внутри интрона и 0 нуклеотидов внутри экзона (Рисунок 1.3). Области внутри экзона не рассматривались, чтобы уменьшить процент ложных положительных предсказаний, поскольку вероятность нахождения консервативной вторичной структуры по случайным причинам выше в последовательностях с высокой средней консервативностью.

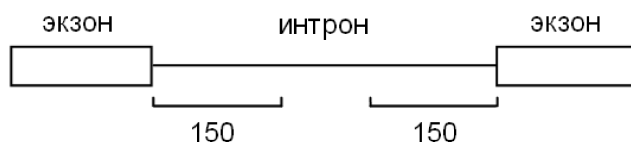


Рис. 1.3. Область поиска консервативных вторичных структур. Рассматривались окружения донорных и акцепторных сайтов сплайсинга протяженностью 150 нуклеотидов внутри интрона.

1.2.1. Предсказанные структуры и уровень ложных положительных предсказаний

Необходимо заметить, что образование вторичной структуры может приводить к выветливанию интрона только в том случае, если структура обладает достаточно большой энергией, чтобы удерживать концы интрона вместе. Выигрыш свободной энергии от спаривания нуклеотидов должен превосходить энергетические затраты на формирование петли [55]. Для гена *CFTTR* показано, что свободной энергии порядка -15 ккал/моль достаточно для формирования вторичной структуры, способной индуцировать пропуск экзона [56]. Эта величина свободной энергии соответствует спирали РНК длиной 8 нуклеотидов [57]. Таким образом, длина спирали 9 нуклеотидов с максимум одной GT парой является более чем достаточной.

Для предсказания вторичных структур был применен метод хэширования (см. Материалы и методы). Мы требовали присутствия хотя бы двух GC пар и консервативности затравки не менее, чем в 9 из 12 видов млекопитающих. При этом консервативность означала попарное сходство (с тремя или менее нуклеотидными заменами) затравок между всеми видами млекопитающих.

С использованием указанных ограничений был получен набор из 211 структур (Таблица 1.3). Для оценки уровня ложных положительных предсказаний применялся контрольный опыт, в котором каждой донорной части интрона сопоставлялась акцепторная часть чужого интрона (см. Материалы и методы). Уровень ложных положительных предсказаний составил $28 \pm 2\%$ (среднее $\pm$ стандартное отклонение в 1000 повторений контрольного опыта) при контроле без ограничения на GC состав или консервативность, $29 \pm 2\%$ при ограничении на GC состав и $36 \pm 2\%$ при ограничении на GC состав и консервативность.

Приведенные оценки уровня ложных положительных предсказаний являются пессимистическими, поскольку вероятность образования вторичных струк-

Таблица 1.3. Количество найденных вторичных структур и оценка уровня ложных положительных предсказаний при контроле без ограничения на GC состав или консервативность (Контроль), при ограничении на GC состав (Контроль+GC) и при ограничении на GC состав и консервативность (Контроль+GC+Cons). Указано среднее значение $\pm$ стандартное отклонение в 1000 повторений контрольного опыта.

Повторы	Структуры	Контроль	Контроль+GC	Контроль+GC+Cons
Не маскированы	211	60.1 $\pm$ 4.2 (28% $\pm$ 2%)	61.6 $\pm$ 4.3 (29% $\pm$ 2%)	76.0 $\pm$ 4.1 (36% $\pm$ 2%)
Маскированы	167	47.4 $\pm$ 3.1 (28% $\pm$ 2%)	43.8 $\pm$ 3.2 (26% $\pm$ 2%)	50.6 $\pm$ 3.0 (30% $\pm$ 2%)

тур повышена в последовательностях, которые содержат повторы, даже после перемешивания донорных и акцепторных частей интронов. Поэтому мы осуществили поиск вторичных структур в том же наборе последовательностей, но с маскированными повторами. Абсолютное количество предсказанных вторичных структур уменьшилось (167 структур), как и уровень ложных положительных предсказаний (30 $\pm$ 2% при ограничении на GC состав и консервативность, Таблица 1.3).

Мы повторили ту же процедуру поиска вторичных структур для затравки длиной 10 нуклеотидов, и уровень ложных положительных предсказаний значительно уменьшился: до 11% в среднем. При этом, вероятно, была потеряна часть реальных структур, поскольку известно и экспериментально подтверждено существование более коротких вторичных структур, которые являются функционально значимыми. Так, длина экспериментально подтвержденных структур в генах β -тропомиозин [58] и дистрофин [59] не превышает семи последовательных Уотсон-Криковских пар. Далее мы будем рассматривать только 9-нуклеотидные затравки с одной GT парой.

1.2.2. Пример предсказанной вторичной структуры

На рис. 1.4 приведен пример одного из интронов, в которых были найдены консервативные вторичные структуры. Здесь можно наблюдать типичную для всего набора интронов картину – вторичные структуры погружены в консервативную область, которая постепенно размывается. При этом необходимо отметить, что длина консервативного участка в данном примере (а также во многих других найденных структурах) больше 9 нуклеотидов. Комплементарность также сохраняется на большей протяженности, чем 9 нуклеотидов. Таким образом, для большинства интронов найденные участки вторичной структуры можно продлить.

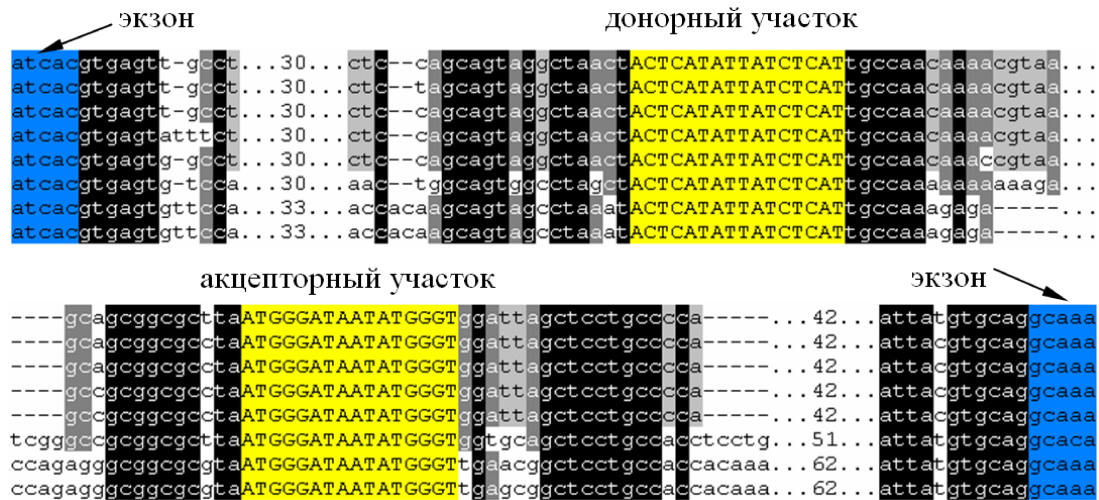


Рис. 1.4. Интрон гена *CG11876*. Синим цветом отмечены границы экзонов, желтым цветом – найденные нами консервативные вторичные структуры, черным цветом – высоко консервативные позиции, серым цветом – позиции средней консервативности.

1.2.3. Регуляция альтернативного сплайсинга

Необходимо отметить, что консервативные вторичные структуры часто встречаются в альтернативно сплайсируемых генах. Рисунок 1.5 показывает, что 33% предсказанных структур соответствуют событиям альтернативного сплайсинга, аннотированным в RefSeq. При этом в контрольном наборе ин-

тронов (без предсказанных структур) того же размера ожидается только 10% структур, соответствующих событиям альтернативного сплайсинга. Кроме того, почти все подтипы альтернативного сплайсинга ассоциированы со вторичными структурами сильнее, чем ожидалось в контрольном наборе интронов.

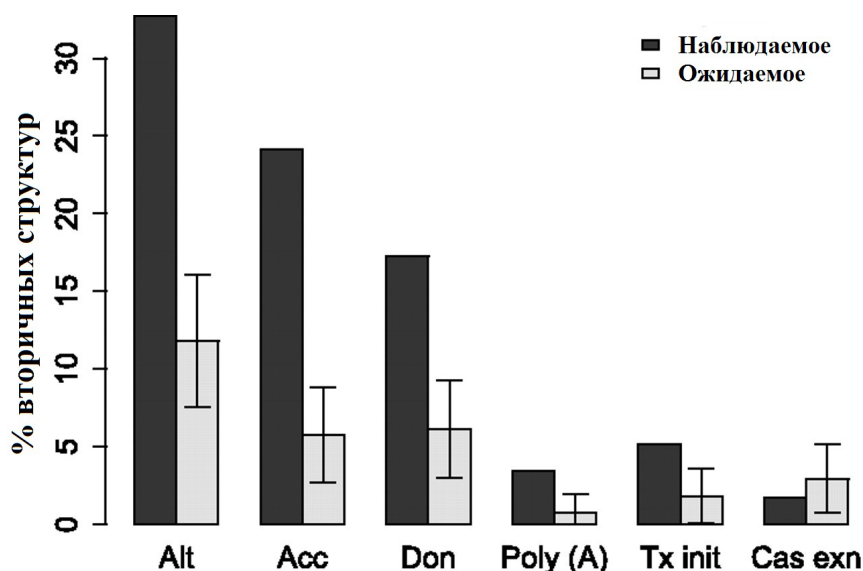


Рис. 1.5. Классы аннотированных событий сплайсинга, ассоциированных со вторичными структурами. Показаны пропорции относительно общего количества предсказанных вторичных структур. Также показаны ожидаемые значения и стандартные отклонения. Alt – альтернативный сплайсинг, Acc – альтернативный акцепторный сайт, Don – альтернативный донорный сайт, Poly(A) – альтернативный сайт полиаденилирования, Tx init – альтернативный сайт инициации транскрипции, Cas expn – интрон, содержащий один или несколько кассетных экзонов.

Можно представить два гипотетических механизма регуляции альтернативного сплайсинга посредством образования вторичных структур РНК: (1) блокирование сайта сплайсинга вторичной структурой, в результате чего сайт сплайсинга оказывается недоступным для сплайсосомы, и используется альтернативный сайт (Рисунок 1.6а); (2) вторичная структура скрепляет концы интрона и сближает сайты сплайсинга, что приводит к преимущественному использованию данного сайта сплайсинга по сравнению с альтернативным (Рису-

нок 1.66).

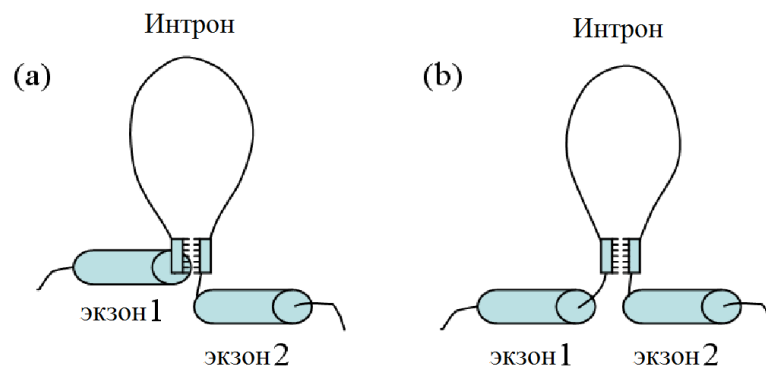


Рис. 1.6. Гипотетические механизмы регуляции альтернативного сплайсинга посредством образования вторичных структур РНК: (а) блокирование сайта сплайсинга; (б) сближение сайтов сплайсинга.

Рассмотрим первый механизм – блокирование сайта сплайсинга. Например, для одного из интронов гена *SLC39A7* при поиске консервативных вторичных структур было замечено следующее. Известно, что этот ген кодирует белок-транспортер, отвечающий за гомеостаз цинка, и альтернативно сплайсируется [60]. Внутри изображенного на рис. 1.7А интрона между экзонами 2 и 3 есть два альтернативных сайта сплайсинга – донорный и акцепторный. Найденная нами вторичная структура закрывает внутренние донорный и акцепторный сайты сплайсинга (Рисунок 1.7В). Можно предположить, что регуляция в данном случае могла бы осуществляться следующим образом: когда участки вторичной структуры спарены, внутренние сайты сплайсинга оказываются недоступным для сплайсосомы, и сплайсинг идет по внешним сайтам (как показано снизу). А если участки не спарены, то сплайсинг может идти по внутренним сайтам (показано сверху).

Второй возможный механизм участия вторичных структур в регуляции альтернативного сплайсинга – сближение сайтов сплайсинга. Было решено проверить, не присутствуют ли в альтернативно сплайсируемых генах такие вторичные структуры, которые могли бы регулировать выбор пути сплайсинга по-

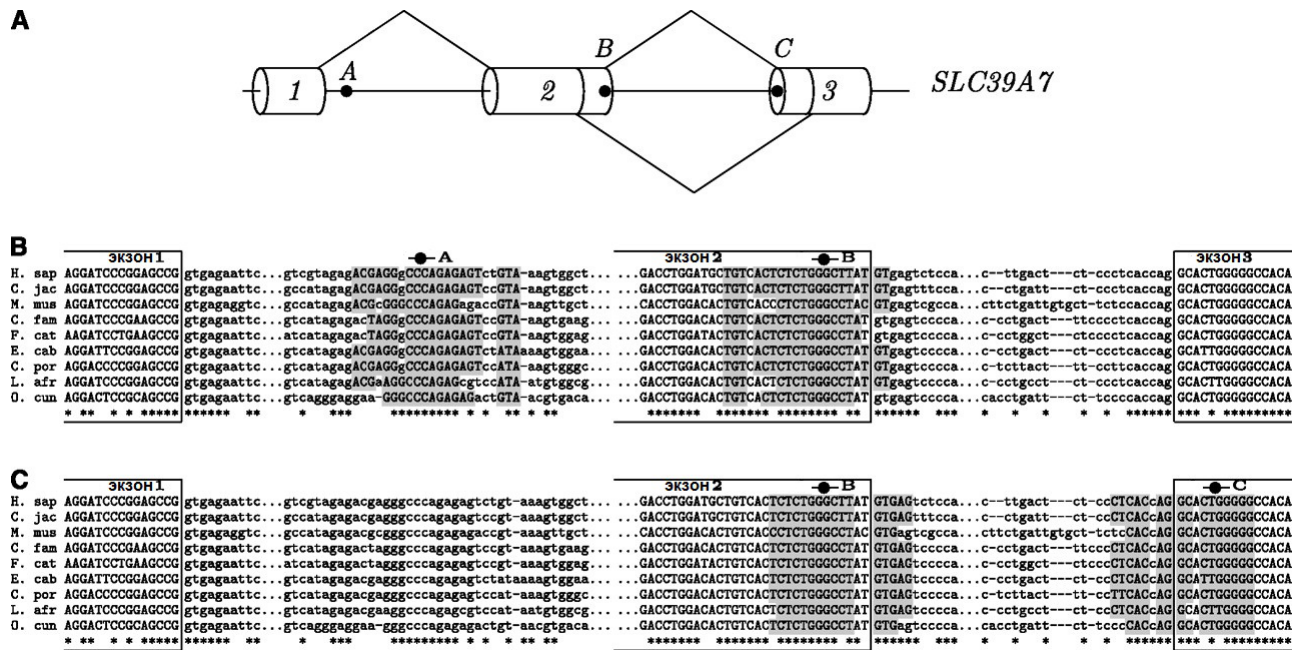


Рис. 1.7. (А) Консервативные вторичные структуры в гене *SLC39A7*. Черными кругами показаны комплементарные участки, параллельными линиями – альтернативные сайты сплайсинга. Сверху и снизу рисунка показаны альтернативные пути сплайсинга, переключение которых, предположительно, регулируется образованием вторичной структуры. (В,С) Множественные выравнивания окружений сайтов сплайсинга в гене *SLC39A7* (консервативные вторичные структуры выделены серым цветом).

добно механизму, рассмотренному выше (Рисунок 1.7). Однако в данном случае мы рассматривали структуры, которые не закрывают сайт сплайсинга, а способствуют сближению разных пар сайтов сплайсинга. Тогда выбор пути сплайсинга в альтернативно сплайсируемом гене будет зависеть от формирования определенной вторичной структуры.

Такие структуры действительно были найдены. На рисунке 1.8А представлен пример одного из этих генов – гена *SF1*. Этот ген кодирует фактор сплайсинга 1 и имеет сложный паттерн альтернативного сплайсинга, а его транскрипт содержит много вторичных структур. Так, вблизи донорных сайтов экзонов 3 и 9 были найдены консервативные участки (С и Е, соответственно), комплементарные одному и тому же участку (F) вблизи акцепторного сайта сплайсинга экзона 10 (Рисунок 1.8В,С). Необходимо заметить, что все структуры длиннее

12 нуклеотидов и высоко консервативны. Предполагается, что при спаривании участков E-F экзоны 4-9 будут преимущественно включаться в процессированный транскрипт, а при спаривании C-F они включаться не будут. При этом важно, что участки C и E не могут одновременно связываться с участком F, и выбор пути сплайсинга будет зависеть от того, какой участок (C или E) взаимодействует с участком F в данный момент.

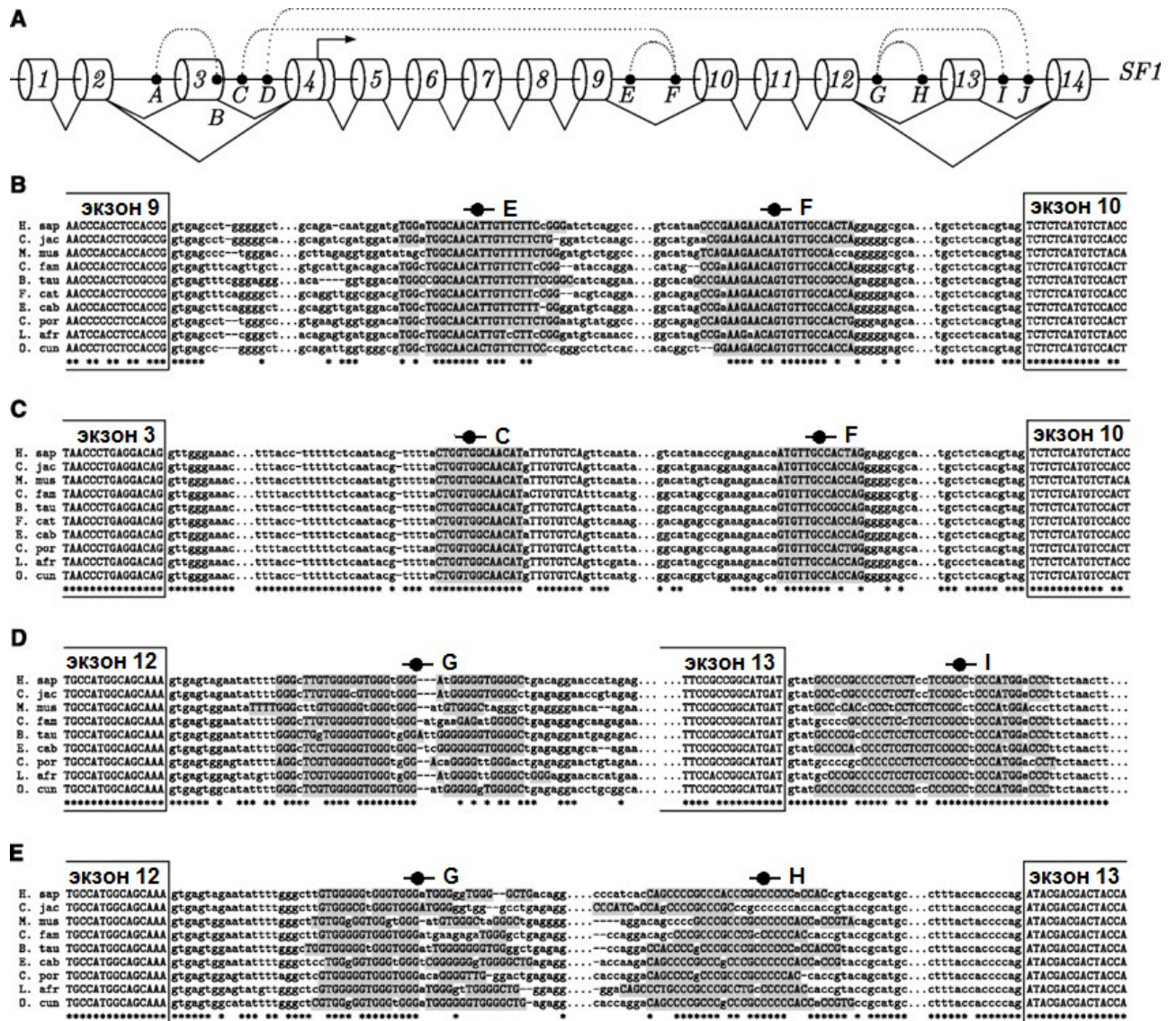


Рис. 1.8. (А) Консервативные вторичные структуры в гене *SF1*. Обозначения как на рис. 1.7А. (В-Е) Множественные выравнивания окружений разных сайтов сплайсинга в гене *SF1* (консервативные вторичные структуры выделены серым цветом).

Предполагается, что спаривание участков C-F приводит к пропуску эк-

зонов 4-9 (Рисунок 1.8). Мы проанализировали публично доступные в базе NCBI Sequence Read Archive (SRA) данные по секвенированию транскриптома (RNA-seq) в клетках лимфатического узла (ERX011193) и щитовидной железы (ERX011194). Прочтения были выравнены на стыки экзонов, которые соответствуют известным событиям сплайсинга, с помощью программы Bowtie с двумя разрешенными ошибками [61]. В среднем, на стыки экзонов картировалось 9.0% прочтений. Те из них, которые картировались в несколько мест одновременно, были удалены (0.8%). Как и было предсказано, транскрипты, соответствующие пропуску экзонов 4-9, были найдены в обоих проанализированных наборах данных.

Кроме того, мы предполагаем, что интрон между экзонами 9 и 10 не вырезается, если участки E и F не спарены. Это предположение косвенно подтверждается тем, что данный интрон содержит стоп-кодон, и его удерживание привело бы к деградации мРНК гена *SF1*. Гипотеза была проверена экспериментально группой Петра Рубцова из Института молекулярной биологии им. В.А. Энгельгардта РАН [52]. Было показано, что нарушение комплементарности участков E и F приводит к вырезанию более длинного интрона из-за преимущественного использования внешнего акцепторного сайта сплайсинга вместо внутреннего акцепторного сайта. Таким образом, можно заключить, что комплементарность участков E-F критически необходима для правильного вырезания интрона.

1.3. Выводы к Главе 1

1. Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, с помощью хэширования.
2. У млекопитающих идентифицировано несколько сотен генов, содержащих

потенциальные консервативные вторичные структуры.

3. Разработан метод определения уровня ложных положительных предсказаний. По пессимистичной оценке этот уровень составляет около 30%.
4. Консервативные вторичные структуры встречаются в альтернативно сплайсируемых генах чаще, чем ожидается.
5. Вторичные структуры могут осуществлять регуляцию альтернативного сплайсинга.
6. Образование вторичной структуры необходимо для правильного вырезания интрона в гене *SF1*.

Глава 2

Регуляция сплайсинга с помощью белковых факторов

2.1. Материалы и методы

Основываясь на данных, полученных с помощью метода iCLIP, совмещенного с высокопроизводительным секвенированием, мы построили точную полногеномную карту позиций связывания белка hnRNPL с РНК в клетках человека (клеточная линия HeLa). Экспериментальные данные были предоставлены группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия (Prof. Dr. Albrecht Bindereif, Justus-Liebig-University-Giessen). Было выполнено три независимых эксперимента и три контрольных эксперимента (без обогащения антителами на белок hnRNPL) с последующим секвенированием на секвенаторе Illumina GAIIx, и в результате получены непарные прочтения длиной 50 нуклеотидов. Экспериментальные процедуры и выравнивание прочтений выполнялись как описано в работе [28].

2.1.1. Обобщение данных о позициях связывания белка hnRNPL в трех независимых экспериментах

Все позиции связывания на аутосомах и хромосоме X, полученные в трех независимых экспериментах, были объединены (всего 1109962 позиций, рис. 2.1A). Чтобы проверить, воспроизводимы ли эксперименты, мы взяли каждую позицию связывания из одного эксперимента (позиция A) и проверили, можно ли найти в 30 или 5 нуклеотидах от нее позицию связывания хотя бы в одном из двух других экспериментов. Такие позиции связывания называли позитивными. Позиции связывания были затем сгруппированы по количеству соответствующим

A

	Прочтения (Solexa 50 осн.)	Позиции связывания
hnRNPL-IP_1	13,327,397	590,960
hnRNPL-IP_2	1,220,636	259,045
hnRNPL-IP_3	1,210,822	303,483
FLAG-IP_1	469,975	18,857
FLAG-IP_2	8,571	1,092
FLAG-IP_3	78,225	9,279

1,109,962 позиции связывания hnRNPL
↓ *clustering*
786,192 кластера позиций связывания
↓ *filtering*
415,606 кластера позиций связывания
(содержат 622,798 позиций связывания)

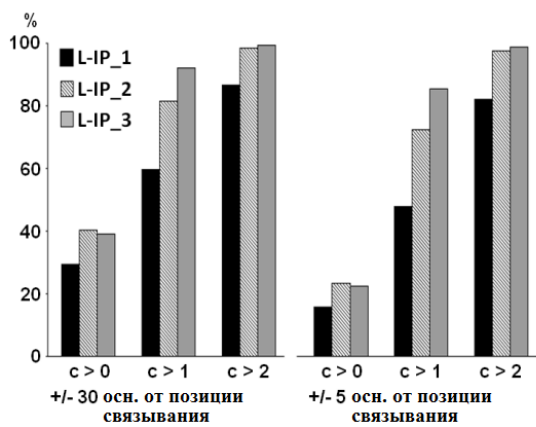
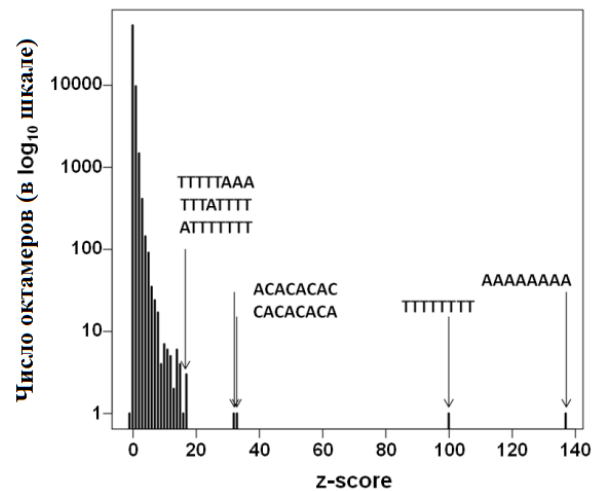
B**C**

Рис. 2.1. (A) Схема обобщения данных, полученных в трех независимых экспериментах (hnRNPL-IP_1, _2 и _3) и в трех контрольных экспериментах (FLAG-IP_1, _2 и _3). (B) Воспроизводимость трех независимых экспериментов. (C) Анализ позиций связывания белка hnRNPL, показывающий частоты различных октамеров, найденных в этих позициях, в виде гистограммы Z-значений.

щих им прочтений следующим образом:

1. все позиции ($c > 0$),
2. позиции с хотя бы двумя прочтениями ($c > 1$),
3. позиции с хотя бы тремя прочтениями ($c > 2$).

Рис. 2.1В показывает процент позитивных позиций связывания при использовании порога на расстояние в 30 нуклеотидов или 5 нуклеотидов. Значения в

первом эксперименте с белком hnRNPL (L-IP_1) были ниже, чем в двух других экспериментах (L-IP_2 и L-IP_3), поскольку количество позиций, определенных в этом эксперименте, было примерно в 10 раз выше, чем в двух других. Поэтому многие позиции из первого эксперимента не воспроизвелись в других экспериментах. Тем не менее, высокий процент позитивных позиций (87% при пороге в 30 нуклеотидов; 82% при пороге в 5 нуклеотидов) говорит о хорошей воспроизводимости экспериментов.

Позиции связывания были сгруппированы в один кластер, если расстояние между ними не превышало 30 нуклеотидов. Далее были проанализированы частоты октамеров в последовательностях 785192 полученных кластеров (включая дополнительные 30 нуклеотидов слева и справа от кластера). Используя статистику Z-значений для частот октамеров, мы обнаружили сильное обогащение Т- и А-богатыми последовательностями (Рисунок 2.1С). Присутствие Т-богатых последовательностей, скорее всего, вызвано систематическими ошибками экспериментальной процедуры (лигирование ультрафиолетом химически вероятнее в случае уридина). Поскольку белок hnRNPL часто связывается с РНК в области 3' UTR (см. Результаты и обсуждение), и плотность позиций связывания возрастает вблизи сайтов полиаденилирования, прочтения, содержащие протяженные А-последовательности, скорее всего, неправильно выравниваются, что приводит к обогащению А-богатыми октамерами.

Известно, что белок hnRNPL преимущественно связывается с СА-повторами [62]. Анализ частот октамеров показал, что СА-богатые мотивы, (СА) $\times$ 4 и (АС) $\times$ 4, встречаются наиболее часто в позициях связывания hnRNPL, но для эффективного функционального анализа белка hnRNPL требуется фильтрация позиций связывания.

Основываясь на результатах работы [62], где были экспериментально (по технологии SELEX) определены последовательности сайтов связывания белка hnRNPL, мы вычислили частоту встречаемости каждого тетрамера в этих сай-

тах связывания. Для каждого кластера позиций связывания был определен индекс качества как сумма частот всех тетрамеров в данном кластере, деленная на длину кластера. Для дальнейшего анализа было отобрано 622798 позиций связывания из 415606 кластеров ($\sim 53\%$ от общего числа кластеров) со значениями индекса качества более 0.06. Большинство отброшенных кластеров содержало Т- и А-богатые последовательности.

2.1.2. События альтернативного сплайсинга, найденные с помощью анализа на микрочипах в клетках HeLa

Группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия (Prof. Dr. Albrecht Bindereif, Justus-Liebig-University-Giessen) был выполнен следующий эксперимент. Из клеток HeLa была выделена тотальная РНК после нокдауна hnRNPL (люцифераза GL2 была использована в качестве контроля) и обработки циклогексимидом. Все 6 образцов (по три биологических реплики для нокдауна hnRNPL и контроля) были подвергнуты стандартному протоколу Affymetrix [GeneChip Whole Transcript (WT) Sense Target Labeling Assay Manual]. Чтобы определить различия в экспрессии экзонов между двумя группами образцов, был использован микрочип GeneChip Human Exon 1.0 Array (<http://www.affymetrix.com>). Нормализованные интенсивности ($\log_2$ значения) экспрессированных проб были отобраны для дальнейшего анализа [63]. Пробы были приписаны к экзонам белок-кодирующих генов в соответствии с аннотацией GENCODE V4. Было выбрано 8376 генов, содержащих более трех экзонов, и с пробами как минимум в 50% их экзонов.

Уровни экспрессии генов были вычислены одновременно как медианы (geneMean) и средние (geneMedian) от интенсивности сигналов всех экзонов, усредненных по всем биологическим репликам в группе образцов. Дифференциальная экспрессия каждого гена между группами образцов (нокдаун hnRNPL — люци-

фераза) была вычислена по значениям geneMean ($\text{diff}_{\text{geneMean}}$) и geneMedian ($\text{diff}_{\text{geneMedian}}$). Для анализа альтернативного сплайсинга были отбраны внутренние экзоны 5828 генов без существенных изменений уровня экспрессии (значения $\text{diff}_{\text{geneMean}}$ и $\text{diff}_{\text{geneMedian}}$ находятся в пределах от -0.25 до 0.25).

Дифференциальная экспрессия каждого экзона ($\text{diff}_{\text{exon}}$) между группами образцов (нокдаун hnRNPL — люцифераза) была вычислена по средним значениям сигнала экспрессии экзона в трех биологических репликах. Уровень включения экзона (EI) был вычислен тремя методами:

1. EI_{mean} : $\text{diff}_{\text{exon}} - \text{diff}_{\text{geneMean}}$,
2. $\text{EI}_{\text{median}}$: $\text{diff}_{\text{exon}} - \text{diff}_{\text{geneMedian}}$,
3. $\text{EI}_{\text{Zscore}}$: Z-значение каждого $\text{diff}_{\text{exon}}$ по отношению ко всем экзонам в гене.

Было выделено два типа событий альтернативного сплайсинга, регулируемых hnRNPL:

1. Активируемые белком hnRNPL: уровень включения экзона уменьшается после нокдауна hnRNPL. Было отобрано 890 экзонов, удовлетворяющих следующему критерию: $\text{EI}_{\text{mean}} \leq -0.75$ И $\text{EI}_{\text{median}} \leq -0.75$ И $\text{EI}_{\text{Zscore}} \leq -2$.
2. Репрессируемые белком hnRNPL: уровень включения экзона увеличивается после нокдауна hnRNPL. Было отобрано 574 экзона, удовлетворяющих следующему критерию: $\text{EI}_{\text{mean}} \geq 0.75$ И $\text{EI}_{\text{median}} \geq 0.75$ И $\text{EI}_{\text{Zscore}} \geq 2$.

Также рассматривалась контрольная группа экзонов, состоящая из всех внутренних экзонов 5828 генов без существенных изменений уровня экспрессии, которые не подходят под критерии (1) и (2).

2.2. Результаты и обсуждение

2.2.1. Консенсусная последовательность позиций связывания hnRNPL богата СА-повторами

Чтобы получить полногеномную карту сайтов связывания hnRNPL в клетках HeLa, группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия было выполнено три независимых эксперимента по технологии CLIP, совмещенной с высокопроизводительным секвенированием. Прочтения из трех экспериментов были объединены и проанализированы, и в результате было аннотировано ~ 1.1 млн. позиций связывания hnRNPL на хромосомах 1-22 и X (см. Материалы и методы).

Анализ содержания пентамеров в последовательностях этих ~ 1.1 млн. позиций связывания был выполнен, как описано в работе [64]. Рассматривались участки последовательности от -30 до -10 и от +10 до +30 по отношению к позиции связывания. В этих участках наблюдалось значительное обогащение СА-повторами и СА-богатыми мотивами (Рисунок 2.2А): СА-повторы (ACACA, CACAC) были представлены чаще ожидаемого на 110-140%, а СА-богатые мотивы (ACAT, TACA) — на 80-95%. Такой паттерн обогащения хорошо согласуется с уже известными свойствами связывания hnRNPL, изученными ранее *in vitro* методом SELEX в работе [62]. В контрольном эксперименте подобного обогащения не наблюдается.

2.2.2. Позиция связывания hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг

Позиции связывания hnRNPL (~ 1.1 млн.) были кластеризованы и профильтрованы, и для дальнейшего функционального анализа было отобрано 622789 позиций связывания (см. Материалы и методы). Распределение таких по-

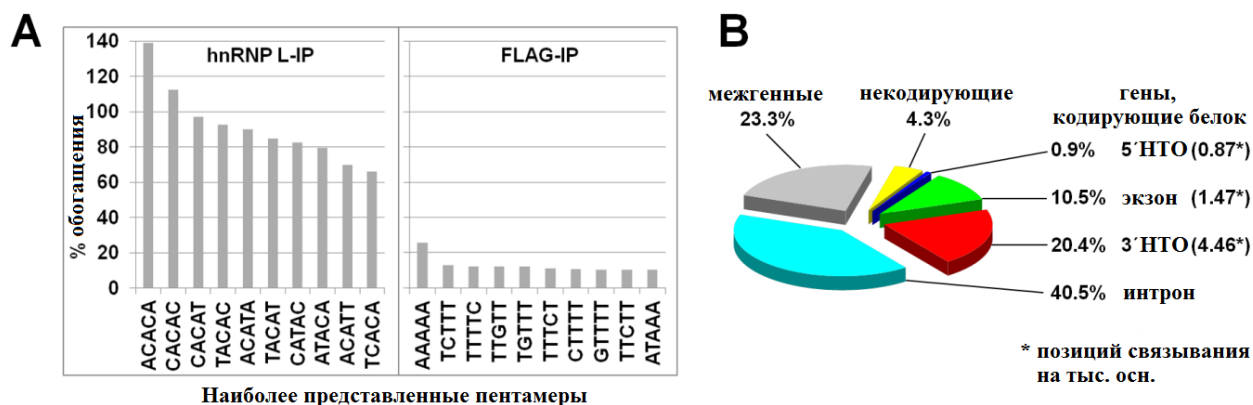


Рис. 2.2. (А) Десять наиболее распространенных пентамеров вблизи позиций связывания hnRNPL (слева) и контрольного FLAG (справа). (В) Распределение позиций связывания hnRNPL в межгенниках, кодирующих и некодирующих генах.

зиций связывания, основанное на аннотации генов GENCODE V4, представлено на рис. 2.2В: большая часть позиций связывания (72.4%) попадает в белок-кодирующие гены, особенно в их интроны (40.5%); в зрелые мРНК попадает меньше позиций связывания (31.9%). Такое большое количество позиций связывания в интронах особенно интересно, учитывая недолгое время жизни интронов, и предполагает существование регуляторного механизма, опосредованного связыванием hnRNPL с интронами. Внутри зрелых транскриптов мРНК, участки 3'НТО демонстрируют гораздо более высокую плотность позиций связывания, чем белок-кодирующие экзоны (4.46 против 1.47 позиций связывания на тыс. нуклеотидов), что хорошо согласуется с известными функциями hnRNPL в области 3'НТО, такими как альтернативное полиаденилирование.

Чтобы исследовать паттерны связывания hnRNPL вблизи сайтов сплайсинга, мы отобрали внутренние экзоны всех аннотированных в GENCODE V4 транскриптов белок-кодирующих генов. Их альтернативные и константные 3'- и 5'-сайты сплайсинга рассматривались отдельно. Количество позиций связывания hnRNPL, найденных вблизи альтернативных и константных 3'-сайтов сплайсинга (интервал от -200 до +75 нуклеотидов, всего 60959 и 143296 альтернативных и константных сайтов сплайсинга, соответственно) и 5'-сайтов

сплайсинга (интервал от -75 до +200 нуклеотидов, всего 57898 и 146357 альтернативных и константных сайтов сплайсинга, соответственно) представлено на рис. 2.3А, верхняя (альтернативные сайты) и средняя (константные сайты) панели.

Для удобства сравнения, количество позиций связывания hnRNPL было нормализовано на отношение количества альтернативных сайтов сплайсинга к константным ($60959 / 143296 = 0.396$ для 3'-сайтов сплайсинга; $57898 / 146357 = 0.425$ для 5'-сайтов сплайсинга) и изображено в виде сглаженных линий (Рисунок 2.3А, нижняя панель).

Наблюдается значительно более высокая плотность позиций связывания hnRNPL вблизи альтернативных 5'-сайтов сплайсинга, как в экзонной (интервал от -30 до 0), так и в интронной (от 0 до +70) части. Кроме того, hnRNPL гораздо чаще связывается вблизи альтернативных 3'-сайтов сплайсинга (около позиции -20). Это наблюдение означает, что hnRNPL может регулировать эффективность использования сайтов сплайсинга посредством связывания вблизи 5'-сайтов сплайсинга, взаимодействуя с интронными или экзонными элементами, а также посредством связывания с полипиримидиновым трактом вблизи 3'-сайтов сплайсинга.

Далее, мы исследовали особенности связывания hnRNPL вблизи экзонамишеней hnRNPL, детектированных с помощью анализа на микрочипах. Экспериментальные данные были предоставлены группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия. Белок hnRNPL был подвергнут нокдауну с помощью РНК-интерференции в клетках HeLa, с последующей обработкой клеток циклогексимидом, чтобы подавить антисмысловой распад мРНК (NMD, или 'nonsense-mediated mRNA decay'). В результате анализа данных (см. Материалы и методы), полученных с помощью микрочипа, было отобрано 890 экзона с пониженным уровнем включения после нокдауна на hnRNPL (hnRNPL действует как активатор) и 574 экзона с повышенным

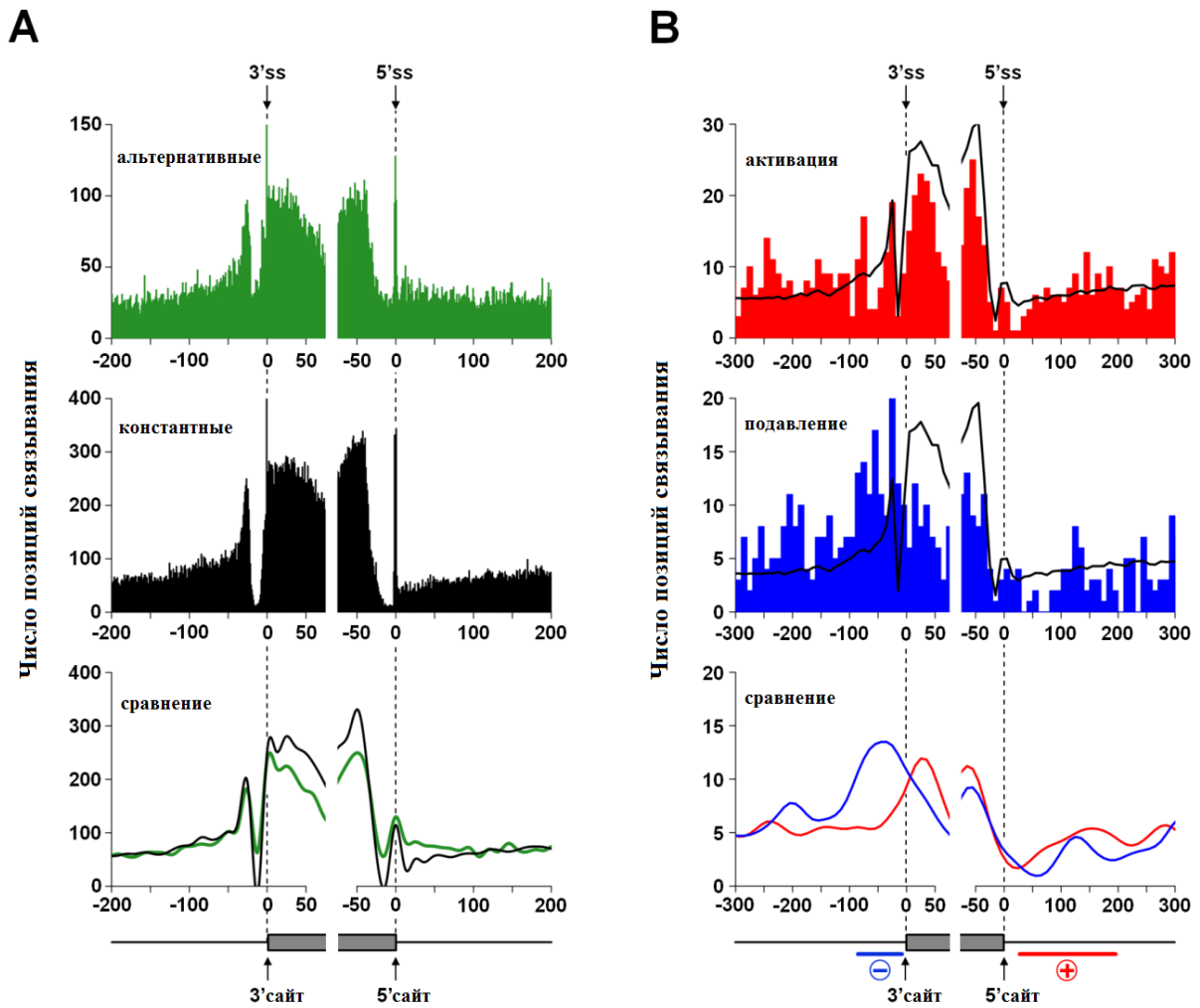


Рис. 2.3. (А) Количество позиций связывания hnRNPL вблизи 3'-сайтов сплайсинга (интервал от -200 до +75) и 5'-сайтов сплайсинга (интервал от -75 до +200) белок-кодирующих генов, отдельно для альтернативных и константных сайтов сплайсинга (верхняя и средняя панели, соответственно). На нижней панели, альтернативные (зеленые) и константные (черные) сайты сплайсинга сравниваются после нормализации и сглаживания. (В) Количество позиций связывания hnRNPL вблизи 3'-сайтов сплайсинга (интервал от -300 до +75) и 5'-сайтов сплайсинга (интервал от -75 до +300) отдельно для экзонов, включение которых активируется или репрессируется (верхняя и средняя панели, соответственно). Контрольный набор экзонов, не регулируемых белком hnRNPL, показан черной линией. На нижней панели, активируемые (красные) и репрессируемые (синие) экзоны сравниваются после нормализации и сглаживания.

уровнем включения (hnRNPL действует как репрессор).

Основываясь на этих двух наборах данных, мы проанализировали плот-

ность позиций связывания hnRNPL отдельно для L-активируемых и L-репрессируемых экзонов (Рисунок 2.3В; верхняя и средняя панели). Их непосредственное сравнение (нижняя панель рис. 2.3В, нормализовано на соотношение репрессированных экзонов к активированным, $574 / 890 = 0.645$) показало, что наблюдаются значительные различия паттерна связывания hnRNPL между L-активируемыми и L-репрессируемыми экзонами. Белок hnRNPL репрессирует включение экзона, если связывается в интронной области, в непосредственной близости от 3'-сайта сплайсинга, и, напротив, активирует включение, если связывается в интронной области вблизи 5'-сайта сплайсинга (первые 200 нуклеотидов). Наблюдаемая высокая плотность позиций связывания hnRNPL внутри активируемых экзонов, по сравнению с репрессируемыми, скорее всего связана с их повышенным уровнем экспрессии.

Таким образом, наши наблюдения показывают, что hnRNPL может подавлять включение экзона, связываясь с полипиримидиновым трактом и препятствуя его распознаванию сплайсосомой, а также активировать включение экзона, связываясь с интронными элементами вблизи 5'-сайта сплайсинга (от 50 до 200 нуклеотидов). Позиция связывания hnRNPL по отношению к регулируемому экзону определяет его активаторную или репрессорную функцию. Данный вывод хорошо согласуется с документированными ранее примерами регуляции сплайсинга посредством белка hnRNPL. Так, было показано активаторное действие hnRNPL для человеческого гена *eNOS*, в котором интронный участок СА-повторов активирует включение экзона 13 [25]. Уровень включения альтернативного экзона 6 человеческого гена *MAPK10* также сильно зависит от энхансера, богатого СА-повторами и расположенного в интервале от +61 до +118 [62]. Кроме того, в работе [63] было приведено несколько примеров подавления включения экзонов в результате связывания белка hnRNPL вблизи их 3'-сайтов сплайсинга, действие которого может быть опосредовано элементом сплайсосомы U2AF, как было показано для гена *TJP1* [65].

2.2.3. Белок hnRNPL часто связывается вблизи мишеней микроРНК в области 3'НТО

Недавно было показано, что белок hnRNPL может конкурировать с микроРНК *miR-297* и *miR-299* за связывание в области 3'НТО гена *VEGFA* при гипоксии и таким образом регулировать подавляющее действие микроРНК на ген *VEGFA* [66]. Чтобы проанализировать данный эффект на полногеномном уровне, мы использовали консервативные мишени микроРНК в области 3'НТО, предсказанные программой TargetScanHuman 5.1 [67]. Сперва мы определили плотность позиций связывания hnRNPL на интервале от -80 до +80 по отношению к границам мишеней микроРНК (Рисунок 2.4А), чтобы продемонстрировать более высокую плотность позиций hnRNPL вблизи мишеней микроРНК.

Затем, среди 3'НТО белок-кодирующих генов, аннотированных в GENCODE V4, было отобрано 5062 3'НТО, содержащих одновременно позиции связывания hnRNPL и предсказанные мишени микроРНК. Для них мы сравнили плотность позиций связывания hnRNPL (количество позиций связывания / длина последовательности) внутри мишеней микроРНК (интервал от -20 до +20 относительно границ мишени) с плотностью вне мишеней, но внутри 3'НТО (Рисунок 2.4В). При этом оказалось, что плотность позиций связывания hnRNPL значительно выше внутри мишеней микроРНК, чем в окружающих 3'НТО (тест Уилкоксона, Р-значение $< 2.2e^{-16}$). Данное наблюдение может означать, что существует глобальный механизм конкуренции между hnRNPL и микроРНК, осуществляющий регуляцию активности транскриптов в цитоплазме.

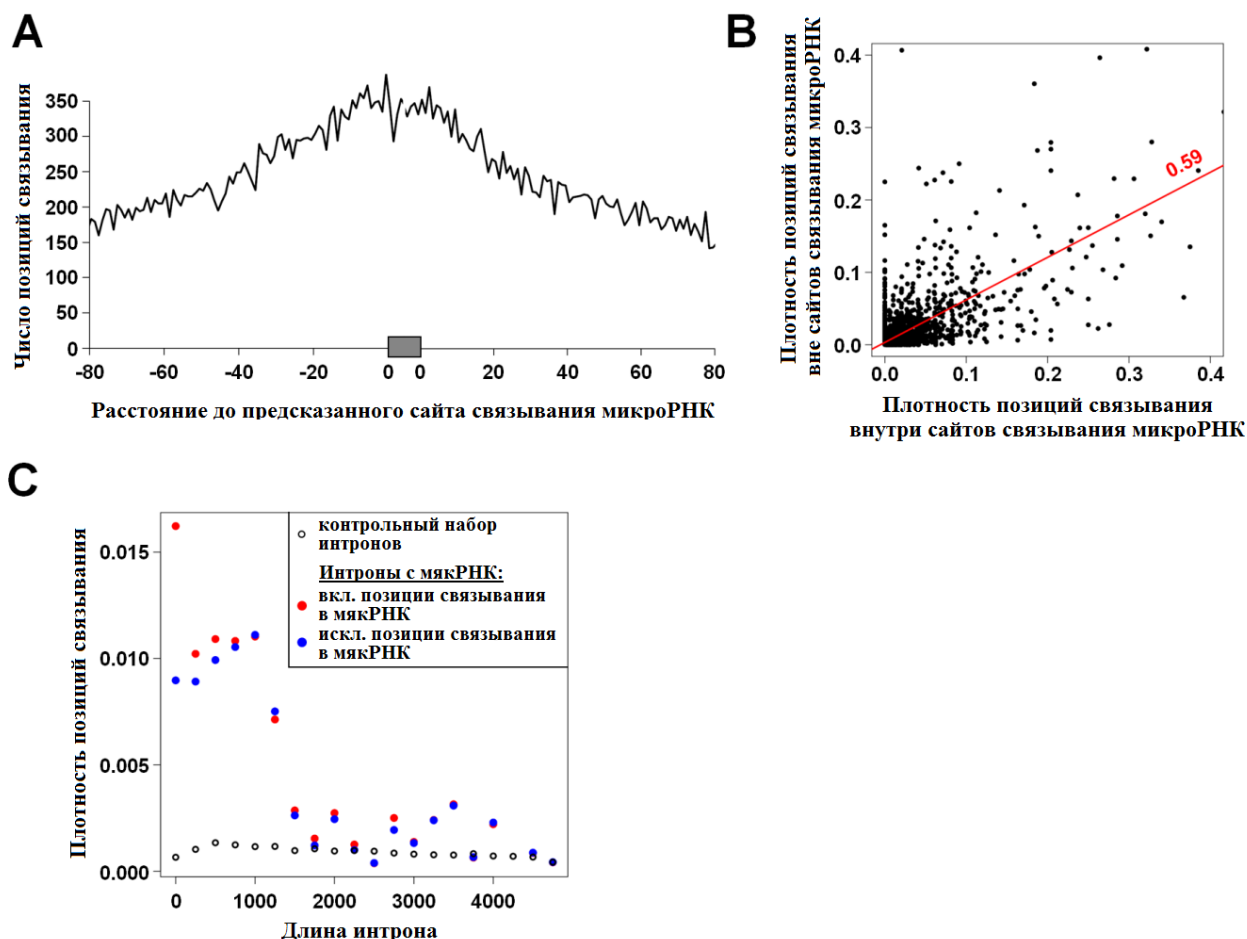


Рис. 2.4. (А) Частота позиций связывания hnRNPL вблизи мишеней микроРНК (интервал от -80 до +80 по отношению к границам мишени). (В) Сравнений плотности позиций связывания hnRNPL между мишенями микроРНК (интервал от -20 до +20) и участками 3'UTRs 5062 белок-кодирующих генов, которые содержат одновременно мишени микроРНК и позиции связывания hnRNPL. (С) Сравнений плотности позиций связывания hnRNPL между мякРНК-содержащими интронами (237 интронов; красные точки — все позиции связывания в интроне; синие точки — все позиции связывания в интроне вне мякРНК) и контрольным набором интронов (8376 интронов; черные незаполненные круги).

2.2.4. Распределение позиций связывания hnRNPL в интронах, содержащих мякРНК

Среди 349 аннотированных мякРНК (малые ядрышковые РНК, или snoRNA), 256 находятся внутри интронов. Большая часть этих интронов (237, или 93%) имеют длину менее 5 тыс. нуклеотидов (166 – 4770 нуклеотидов). Именно для

таких интронов мы проанализировали плотность позиций связывания hnRNPL. 106559 интронов длиной менее 5 тыс. нуклеотидов из 8376 экспрессированных мультиэкзонных белок-кодирующих генов (см. Материалы и методы) были использованы в качестве контроля. Интроны каждой группы были разбиты на 20 категорий в соответствии с их длиной (с интервалом в 250 нуклеотидов: 1-250, 251-500, 501-750, ..., 4750-5000). Средние значения плотности позиций связывания hnRNPL (количество позиций / длина интрона) были вычислены отдельно для (Рисунок 2.4С):

1. всего интрона, содержащего мякРНК (красные точки),
2. области интрона вне мякРНК (синие точки),
3. всего интрона, не содержащего мякРНК (черные незаполненные круги).

Оказалось, что плотность позиций связывания hnRNPL в интронах, содержащих мякРНК, значительно выше. Данное наблюдение может свидетельствовать об участии белка hnRNPL в метаболизме мякРНК. Оно было подтверждено экспериментально группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия [68].

2.3. Выводы к Главе 2

1. Построена полногеномная карта позиций связывания белка hnRNPL с РНК в клетках человека HeLa.
2. Сайты связывания hnRNPL обогащены СА-повторами и СА-богатыми мотивами.
3. Распределение позиций связывания hnRNPL вокруг 5'- и 3'-сайтов сплайсинга различается между альтернативными и константными экзонами, и между L-активируемыми и L-репрессируемыми экзонами.

4. Позиция связывания белка hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг.
5. Белок hnRNPL часто связывается вблизи мишеней микроРНК в области 3'UTR.
6. Плотность позиций связывания hnRNPL в интронах, содержащих мякРНК, значительно выше, что может говорить об участии белка hnRNPL в метаболизме мякРНК.

Глава 3

Транс-сплайсинг

3.1. Материалы и методы

3.1.1. Анализ систематических ошибок секвенирования

Данные были получены из базы NCBI Sequence Read Archive (SRA). Данные в этой базе организованы в наборы экспериментов, которые были выполнены в рамках одного исследования в одной и той же лаборатории, предположительно по одному и тому же протоколу. Каждый эксперимент содержит данные, полученные в результате одного или более запусков секвенирования. Мы отобрали для последующего анализа такие эксперименты, общее количество данных в которых было не менее 500 миллионов нуклеотидов, и которые были публично доступны на момент исследования (10.10.2010). В каждом эксперименте мы выбрали несколько запусков секвенирования, чтобы составить для этого эксперимента набор данных из около 1.5 миллиардов нуклеотидов. Исследования с идентификаторами SRP001106, SRP001699, SRP001734, SRP002009 и SRP002881 не были включены в анализ, поскольку размер каждого из их запусков секвенирования превосходил 5 миллиардов нуклеотидов.

Прочтения были выравнены на референсный геном человека (версия hg19) с помощью программы *bowtie* [61]. Количество разрешенных ошибок менялось в зависимости от длины прочтения: одна ошибка для прочтений короче 26 нуклеотидов, две ошибки для интервала 26-50 нуклеотидов, и три ошибки для прочтений длиннее 50 нуклеотидов. Прочтения, полученные на платформах компании Иллюмина (Illumina), были выравнены в нуклеотидном пространстве, а прочтения, полученные на платформах компании Солид (SOLiD), были выравнены в цветовом пространстве. Прочтения, которые выравнивались сразу в несколько

мест, не рассматривались.

Профили покрытия генов были вычислены для каждой позиции гена как общее количество прочтений, покрывающих данную позицию. Для каждого гена был вычислен коэффициент корреляции Пирсона между его профилями покрытия во всех возможных парах экспериментов. Профили покрытия не были предварительно нормализованы на общее количество прочтений в каждом эксперименте, поскольку коэффициент корреляции Пирсона от этого не зависит.

Гены, которые были покрыты прочтениями в менее, чем 90% экспериментов, не рассматривались. Такую фильтрацию прошли 178 из 1074 одноэкзонных белок-кодирующих генов. Коэффициенты корреляции были усреднены по всем этим генам для каждой пары экспериментов, которые были впоследствии кластеризованы по методу UPGMA [69] в среде R [70].

Коэффициент выравнивания прочтений ('read mappability') – это один из возможных источников систематических ошибок секвенирования. Он основан на том наблюдении, что прочтения выравниваются хуже на участки генома с низкой сложностью последовательности. Коэффициент выравнивания прочтений был вычислен следующим образом. Для каждого прочтения длины L , представленной в анализируемом наборе данных, был построен профиль выравнивания путем извлечения L -меров, начинающихся в каждой позиции генома, и их последующего выравнивания на геном с точно теми же параметрами, что и экспериментальные данные.

3.1.2. Поиск химерных РНК

Поиск химерных РНК осуществлялся в трех наборах данных секвенирования транскриптома человека (ткань мозга [71], клеточная линия эритролейкемии K562 [72]), лимфобластоидная клеточная линия GM12878 [73]), а также с помощью базы данных ChimerDB [74]. Парные прочтения были картированы

на референсный геном человека (версия hg18) и на все возможные внутригенные стыки экзонов с помощью программы SOAP [75]. Для дальнейшего анализа мы отобрали химерные пары прочтений, состоящие из прочтений, выравненных на разные хромосомы. Таких пар было найдено 431321 в образце ткани мозга, 907368 в клеточной линии GM12878 и 361487 в клеточной линии K562.

3.1.3. Контрольный набор химерных РНК

Контрольный набор данных был получен с помощью составления искусственных пар для каждого прочтения со случайным прочтением на другой хромосоме. Такой процедуре была подвергнута каждая пара прочтений, состоящая из прочтений, выравненных на одну и ту же хромосому. Полученный набор данных, состоящий из искусственных химерных РНК, сохраняет изначальную плотность генов, поскольку локальная плотность покрытия генома прочтениями не меняется.

3.1.4. Карта взаимодействий хромосом в пространстве

В матрице пространственной близости M , сконструированной для человеческого генома с помощью метода Hi-C [40], каждое значение $m(i, j)$ определяется как количество продуктов лигирования между фрагментами ДНК i и j длиной в 1 миллион нуклеотидов (1Мб) каждый. В работе [40] матрица M была нормализована на покрытие, при этом была получена новая матрица M^* следующим образом. Рассматривались только пары фрагментов, находящихся на разных хромосомах. Ожидаемое количество взаимодействий в каждой паре фрагментов i, j было вычислено путем умножения доли прочтений, соответствующих фрагменту i , на долю прочтений, соответствующих фрагменту j , и на общее количество прочтений. Для каждой пары фрагментов i и j , обогащение $m^*(i, j)$ было вычислено как наблюдаемое количество взаимодействий $m(i, j)$,

деленное на ожидаемое количество взаимодействий. Чтобы улучшить разрешение, была также построена корреляционная матрица C , в которой $c(i, j)$ равно коэффициенту корреляции Пирсона между i -м рядом и j -й колонкой матрицы M^* .

3.1.5. Данные о функциональном состоянии хроматина

Чтобы получить данные о функциональном состоянии хроматина, были использованы данные нескольких исследований. Экспрессионные данные для лимфобластной клеточной линии GM06690 были получены из работы [40]. Данные о модификациях гистонов, метилировании ДНК и чувствительности к ДНКазе были получены из проекта ENCODE для лимфобластной клеточной линии GM12878 [73]. При этом данные о пространственной близости фрагментов генома были сконструированы для лимфобластной клеточной линии GM06990 [40].

Полногеномная аннотация 'состояний хроматина', или биологически значимых комбинаций эпигенетических маркеров, была получена из работы [51]. Данная аннотация включает в себя 51 состояние хроматина, которые были построены на основе 38 разных меток метилирования и ацетилирования гистонов, а также на основе варианта гистона H2AZ, полимеразы II типа и белка CTCF в CD4 T-лимфоцитах человека.

3.1.6. Контрольный набор данных о функциональном состоянии хроматина

Существует два типа хроматина: 'закрытый' хроматин, бедный генами, и 'открытый' хроматин, богатый генами. Известно, что транскрипционно активные и богатые генами участки генома предпочитают располагаться рядом с другими активными участками (даже если последние располагаются на дру-

гой хромосоме) и пользоваться общими транскрипционными факторами и регуляторными элементами, образуя транскрипционные фабрики. Поэтому можно предположить, что подобие эпигенетических свойств в пространственно близких участках генома может объясняться высокой плотностью генов в этих участках.

Для контроля влияния геномных доменов, обогащенных и обедненных генами, мы создали контрольный набор данных следующим образом. В соответствии с аннотацией компартментов открытого/закрытого хроматина, предложенной в работе [40], мы приписали каждый ген к тому компартменту, в котором находится начало кодирующей последовательности данного гена. Имена генов были перемешаны случайным образом внутри одного и того же компартмента, при этом позиции генов оставались нетронутыми. Такая процедура не повлияла на локальную плотность генов, то есть корреляции между пространственной близостью и плотностью генов в парах геномных фрагментов остались неизменными.

3.2. Результаты и обсуждение

3.2.1. Анализ систематических ошибок секвенирования

Используя данные секвенирования генома и транскриптома человека из 117 экспериментов, выполненных в 26 лабораториях по всему миру на разных платформах, мы вычислили профили покрытия генов, которые используются для решения таких популярных задач, как определение уровней экспрессии генов или включения экзонов. Были рассмотрены только одноэкзонные гены с высоким покрытием (178 генов, см. Материалы и методы), чтобы сделать профили покрытия генов сравнимыми между экспериментами по секвенированию мРНК в разных тканях, которые могут производить альтернативные изоформы

в результате сплайсинга, а также между экспериментами по секвенированию генов и транскриптомов.

Кластеризация экспериментов по профилям покрытия генов показана на Рисунке 3.1. Подписи содержат следующую информацию об экспериментах: идентификатор исследования в SRA; идентификатор эксперимента в SRA; краткое название лаборатории; секвенирование генома ("G") или РНК ("R"); платформа ("T" обозначает платформу Иллюмина, "S" – Солид); размер фрагмента, если прочтения парные; размер прочтения; идентификатор образца, популяция, клеточная линия и/или ткань. Кластеризация показала, что эксперименты по секвенированию транскриптома, сделанные в одной и той же лаборатории, имеют очень схожие профили покрытия, даже если были секвенированы такие разные ткани, как мозг и печень. Средний коэффициент корреляции между профилями покрытия одного и того же гена в разных экспериментах в одной и той же лаборатории составляет 0.46 ± 0.14 для секвенирования РНК на платформе Иллюмина (Рисунок 3.2). В то же время, эксперименты по секвенированию транскриптома, сделанные в разных лабораториях, имеют значительно менее похожие профили покрытия, даже если была секвенирована одна и та же ткань. Средний коэффициент корреляции в этом случае составляет 0.27 ± 0.10 (Рисунок 3.3). Мы применили критерий суммы рангов Уилкоксона для проверки того, что средний коэффициент корреляции профилей покрытия выше внутри лаборатории, чем между лабораториями. Поскольку полученное Р-значение оказалось меньше $2.2e^{-16}$, можно говорить о том, что профили покрытия генов действительно более похожи у экспериментов, выполненных в одной и той же лаборатории, чем в разных лабораториях.

Для экспериментов по секвенированию генома такой четкой картины не наблюдается. Некоторые из них кластеризуются по лабораториям, другие – нет. Все эксперименты, выполненные на платформе Солид, собираются в один кластер. При этом эксперименты по секвенированию транскриптома, выполненные

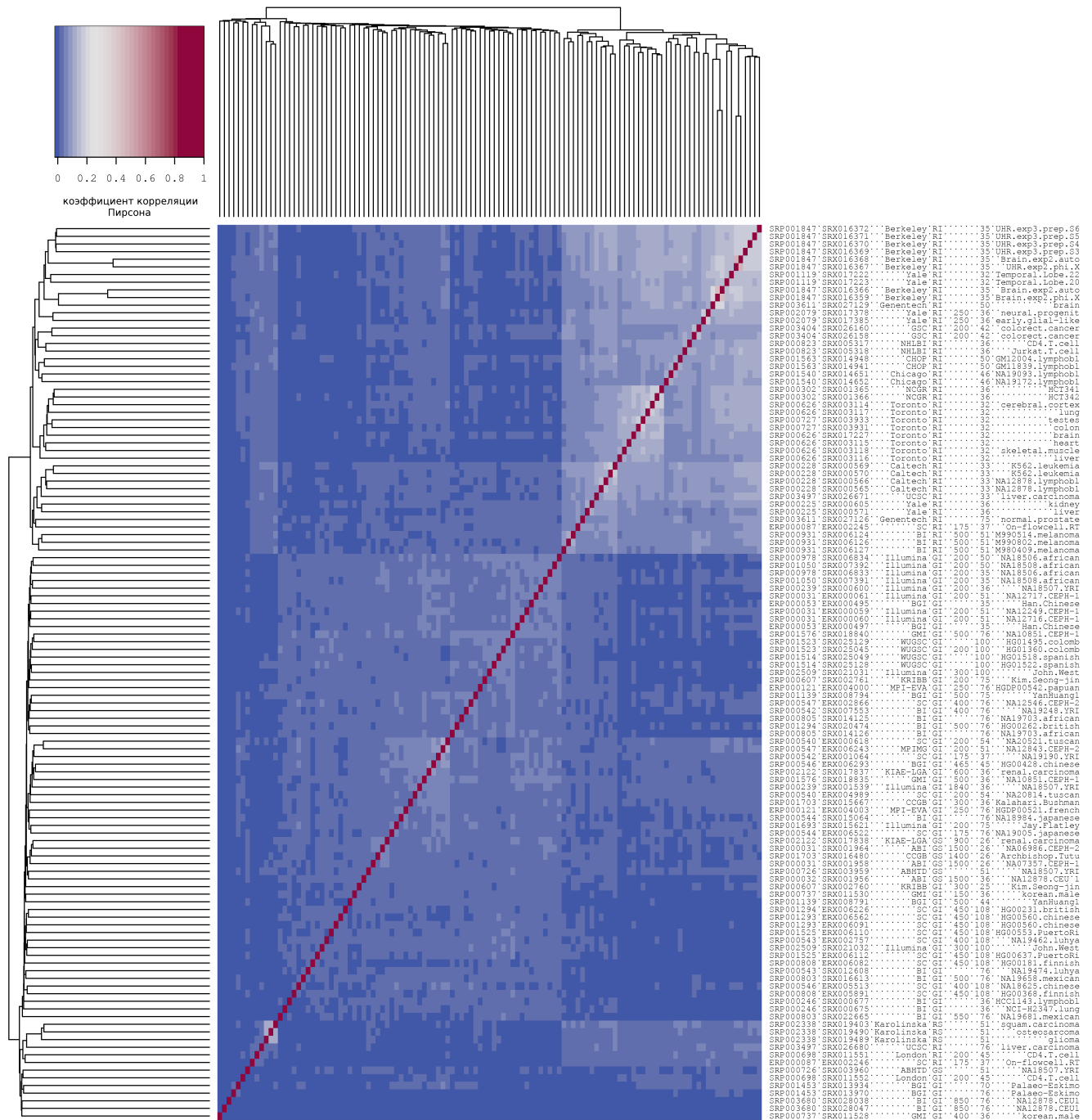


Рис. 3.1. Кластеризация экспериментов по профилям покрытия генов. Цвета отображают значения коэффициента корреляции Пирсона: оттенки синего соответствуют интервалу $(-0.1, 0.2)$, оттенки красного – интервалу $(0.2, 1.0)$. См. описание в тексте.

на платформе Солид, не похожи на все остальные эксперименты.

Наблюдаемая зависимость профилей покрытия генов от лаборатории говорит о том, что результаты секвенирования надо сравнивать с осторожностью. Эта проблема стоит наиболее остро для экспериментов по секвенированию тран-

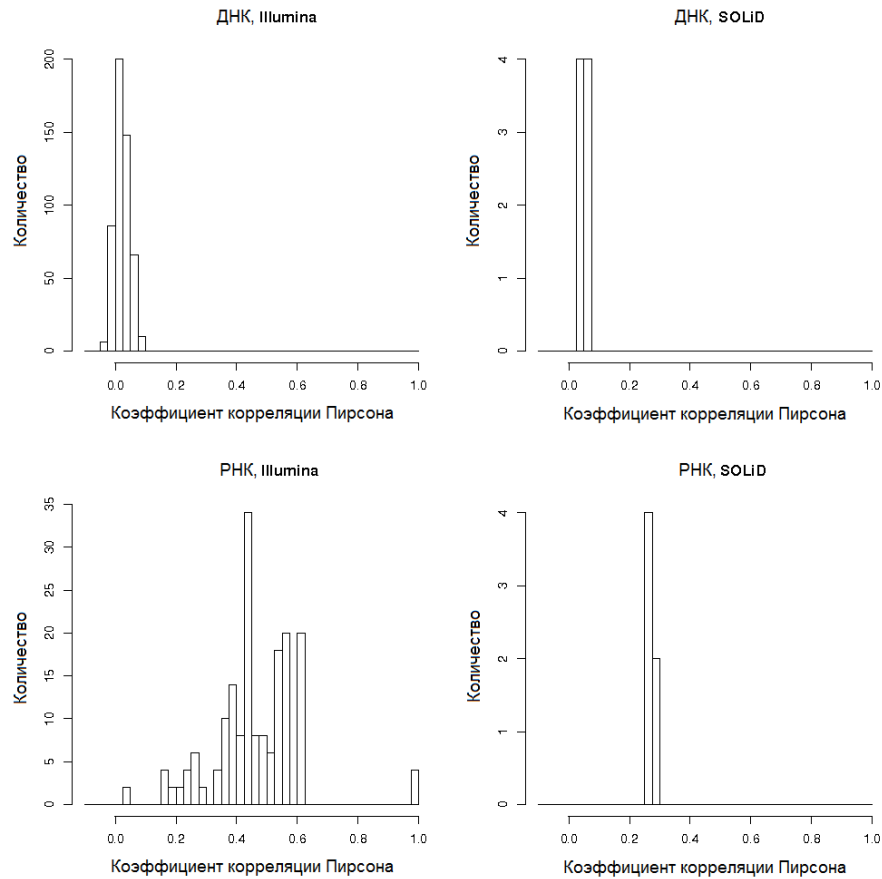


Рис. 3.2. Распределение коэффициентов корреляции Пирсона в одной и той же лаборатории. Коэффициенты корреляции Пирсона были вычислены между всеми возможными парами экспериментов в одной и той же лаборатории, независимо для каждого одноэксонного гена, и впоследствии усреднены по всем таким генам.

скрипта, поскольку целью большинства таких экспериментов является именно сравнение, будь то уровни экспрессии в разных тканях, здоровые и больные индивидуумы, и т.д. Поскольку эксперименты по секвенированию генома не кластеризуются по лабораториям так четко, на них эта систематическая ошибка, скорее всего, не распространяется.

Процедура кластеризации была повторена для нефильтрированного набора одноэксонных генов (1074 гена, см. Материалы и методы), и был получен сходный результат. Таким образом, наши наблюдения не зависят от критерия фильтрации.

Далее были построены линейные модели для профилей покрытия, усред-

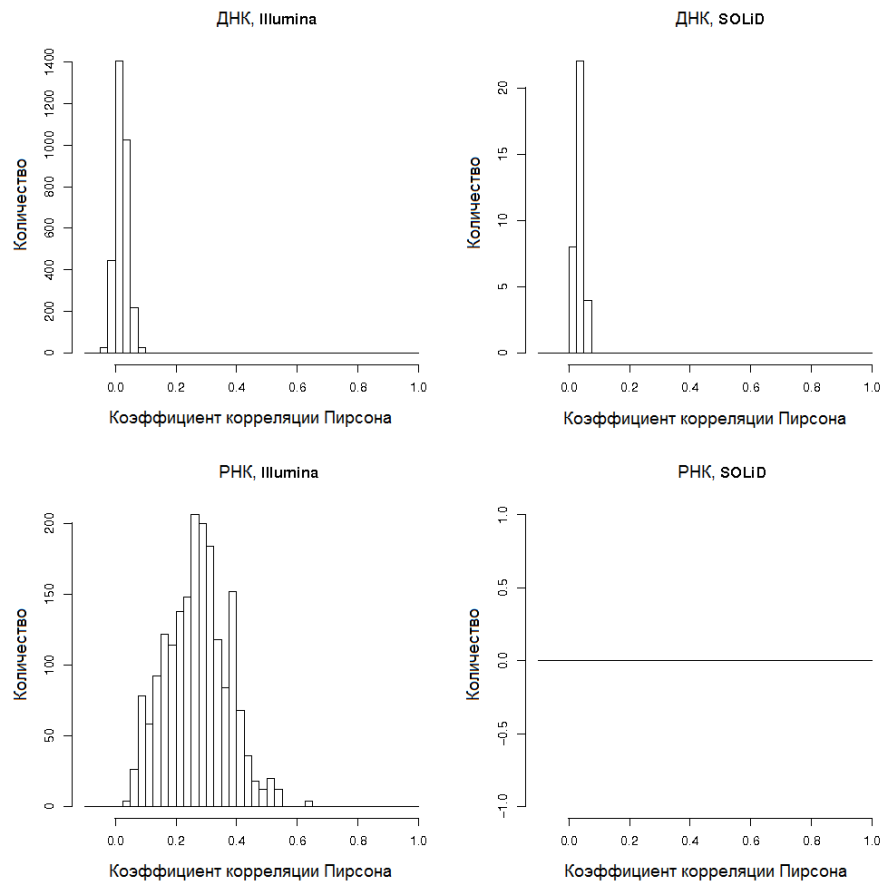


Рис. 3.3. Распределение коэффициентов корреляции Пирсона в разных лабораториях. Коэффициенты корреляции Пирсона были вычислены между всеми возможными парами экспериментов в разных лабораториях, независимо для каждого одноэкзонного гена, и впоследствии усреднены по всем таким генам.

ненных по всем генам в данном эксперименте. Большинство экспериментов по секвенированию транскриптома на платформе Иллюмина демонстрируют систематическую ошибку покрытия, при которой 3'-концы генов покрыты лучше, чем 5'-концы. Эксперименты по секвенированию транскриптома на платформе Солид, напротив, демонстрируют более высокое покрытие 5'-концов генов, чем 3'-концов. Чтобы скорректировать эти систематические ошибки, мы нормализовали все профили покрытия в соответствии с их линейными моделями, и снова кластеризовали эксперименты. Существенных изменений не произошло. Нормализованные эксперименты по секвенированию транскриптома кластеризуются так же хорошо, как ненормализованные (средний коэффициент корреляции

внутри лаборатории 0.43 ± 0.21 , между лабораториями – 0.25 ± 0.14).

Неравномерность покрытия генов от 5'- к 3'-концу была известна ранее и может быть вызвана спецификой протоколов пробоподготовки [76, 77]. Эта систематическая ошибка практически не влияет на исследования, включающие сравнение результатов экспериментов, выполненных в одной и той же лаборатории (средний коэффициент корреляции 0.46 ± 0.14). Однако, сравнение результатов экспериментов, выполненных в разных лабораториях, должно проводиться с большой осторожностью (средний коэффициент корреляции 0.27 ± 0.10).

Также было замечено, что некоторые участки генов покрыты выше, чем другие, и их относительные позиции в генах повторяются внутри лаборатории (Рисунок 3.4). Этот эффект нельзя объяснить длиной прочтения, которая в большинстве случаев одна и та же внутри лаборатории, т.к. эксперименты не кластеризуются по длине прочтения. Мы попытались скорректировать данную систематическую ошибку путем нормализации профилей покрытия на коэффициент выравнивания прочтений ('read mappability') – относительную величину, которая для каждой позиции генома показывает, насколько хорошо прочтения выравниваются на данную позицию (см. Материалы и методы). Однако, данная нормализация также не дала существенных результатов. Некоторые участки генов были по-прежнему покрыты лучше других, причем относительные позиции таких участков были уникальными внутри каждой лаборатории, что позволяет говорить о существовании систематической ошибки неизвестной природы, характерной для экспериментов по секвенированию транскриптома. Одним из возможных объяснений такой неравномерности покрытия может быть вторичная структура РНК [37].

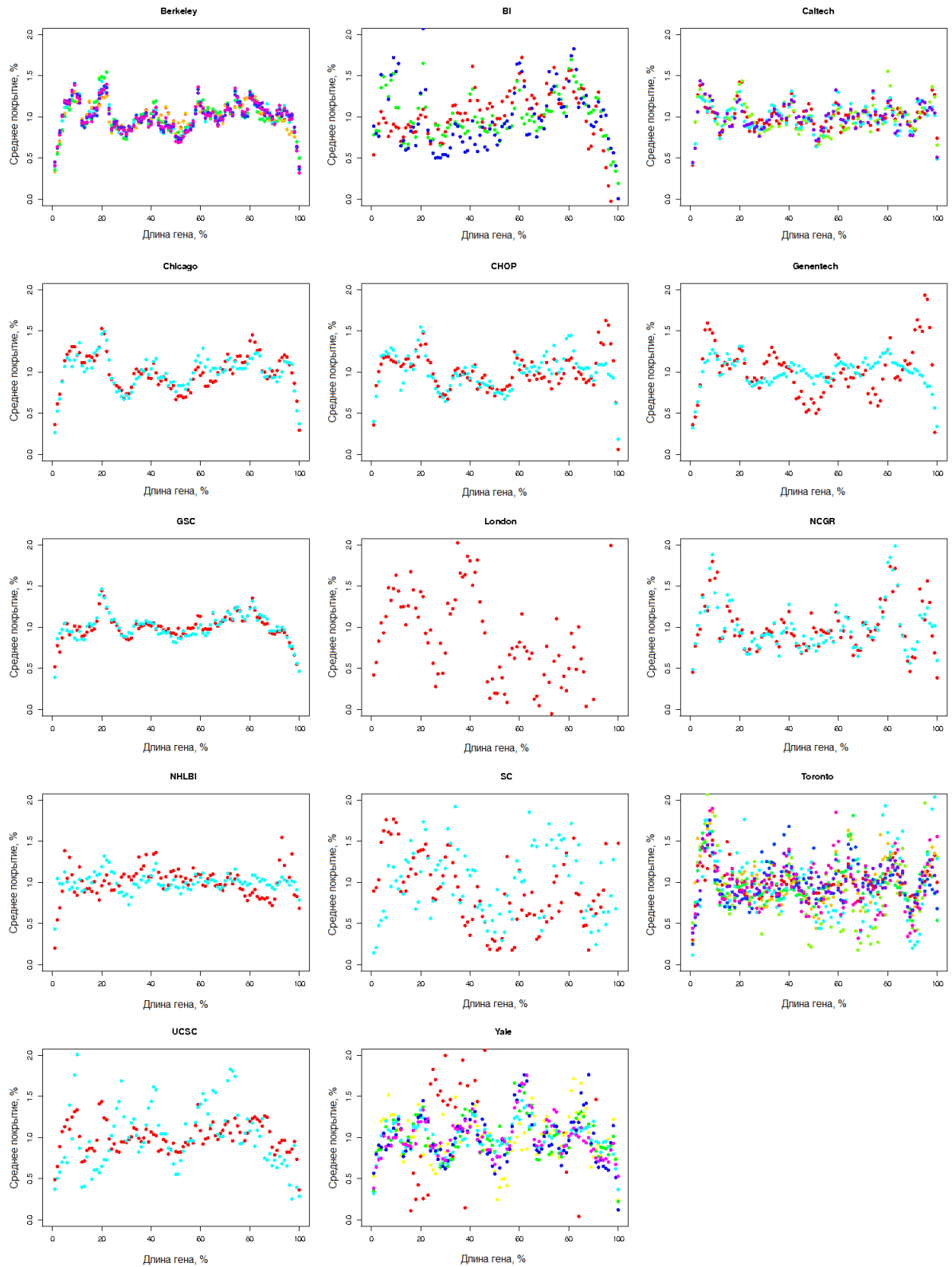


Рис. 3.4. Распределение покрытия одноэкзонных генов по длине. Распределение вычислено после нормализации на неравномерность покрытия от 5'- к 3'-концу. Цветные точки обозначают разные эксперименты, сгруппированные по лабораториям.

3.2.2. Удаление систематических ошибок в данных о пространственной структуре генома человека

Карта контактов участков ДНК в пространстве была построена с помощью методов высокопроизводительного секвенирования [40], и поэтому может быть загрязнена систематическими ошибками секвенирования, происходящими в результате экспериментальной процедуры (а именно, полимеразной цепной реакции) или неправильного выравнивания прочтений. Оба этих типа систематических ошибок происходят наиболее часто между участками генома с высоким уровнем сходства. Если в геноме есть два таких локуса с высоким уровнем сходства, то программы выравнивания прочтений могут ошибиться и выравнивать второе прочтение пары на удаленный локус генома, вместо того, чтобы выравнивать его на тот же самый локус, на который выровнено первое прочтение пары. ДНК-полимераза (фермент, который осуществляет полимеразную цепную реакцию) также может ошибиться и перепутать участки генома, содержащие достаточно протяженные идентичные последовательности [78].

В связи с этим, мы вычислили содержание идентичных последовательностей во взаимодействующих фрагментах ДНК (Рисунок 3.5А). Здесь и далее рассматривались только пары фрагментов, находящиеся на разных хромосомах, чтобы избежать необходимости учета линейного расстояния между фрагментами. Последовательности с высокой идентичностью были определены с помощью программы *blastn* [79], с порогом на идентичность 92%, что соответствует 6 ошибкам на прочтение длиной 75 нуклеотидов. Уровень идентичности вычислялся для каждой пары взаимодействующих фрагментов i и j как сумма длин последовательностей с высокой идентичностью между этими фрагментами, деленная на длину фрагмента.

Значения пространственной близости были условно разделены на 29 интервалов. Аномально высокое содержание идентичных последовательностей на-

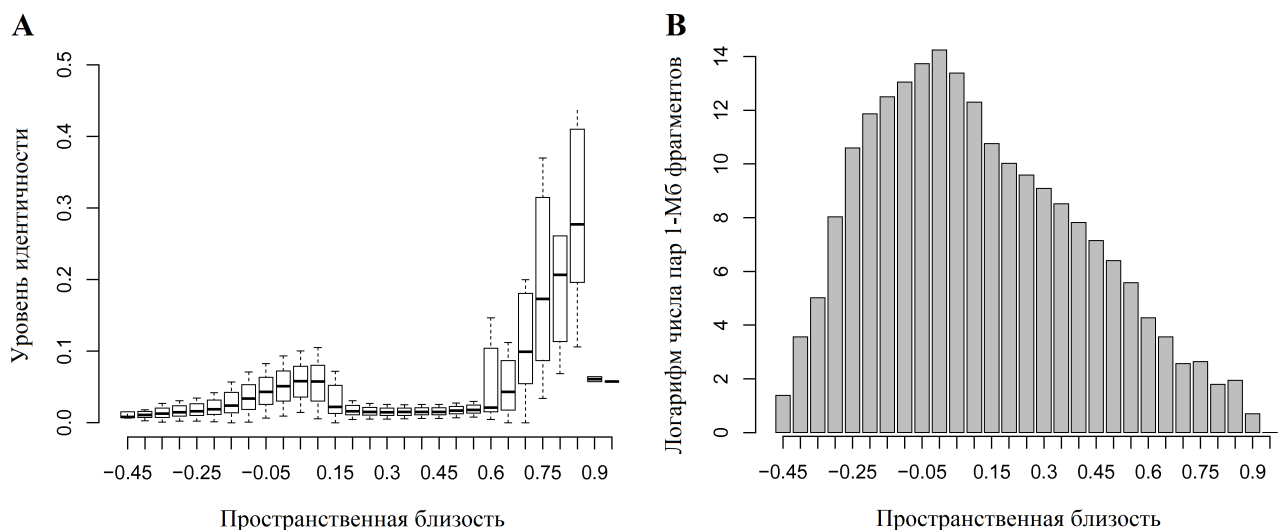


Рис. 3.5. (А) Содержание идентичных последовательностей в 29 рассматриваемых интервалах пространственной близости между 1-Мб фрагментами разных хромосом в полногеномной корреляционной матрице C [40] (см. Материалы и методы). Отрицательные значения пространственной близости соответствуют удаленным фрагментам, положительные значения – сближенным фрагментам. На графике показаны квантили (прямоугольник), медиана (линия в прямоугольнике), минимальное и максимальное значения (линии вне прямоугольника). (В) Гистограмма количества пар взаимодействующих геномных фрагментов, происходящих с разных хромосом в полногеномной матрице пространственной близости C .

блюдается в геномных фрагментах со значениями пространственной близости выше 0.55, указывая на то, что некоторые из этих 'пространственно близких' фрагментов могут в действительности не располагаться близко друг к другу в ядре, а являться результатом систематических ошибок экспериментального протокола секвенирования или выравнивания прочтений. Кроме того, общее количество пар фрагментов в концевых интервалах существенно меньше, чем в центральных интервалах (Рисунок 3.5В), и результаты для таких интервалов имеют низкую статистическую значимость. Поэтому далее мы рассматривали только интервалы с пространственной близостью от -0.3 до 0.55 .

Для фрагментов со значениями пространственной близости около нуля также заметен отчетливый пик содержания идентичных последовательностей. Чтобы понять его источник, мы проанализировали содержание повторов во вза-

имодельствующих фрагментах, используя данные из базы UCSC Genome Browser Database [80] (Рисунок 3.6). Содержание повторов было вычислено как среднее количество нуклеотидов, маскированных программой *RepeatMasker* в двух взаимодействующих 1-Мб фрагментах ДНК. При этом пик около нуля наблюдался в случае особого класса повторов, называемых 'exapted repeats', – консервативных повторов вне экзонов, которые предположительно возникли при участии мобильных элементов [81]). Данное наблюдение означает, что такие повторы могут быть перепредставлены в парах фрагментов со значениями пространственной близости около нуля, однако требуется дополнительное, более детальное исследование.

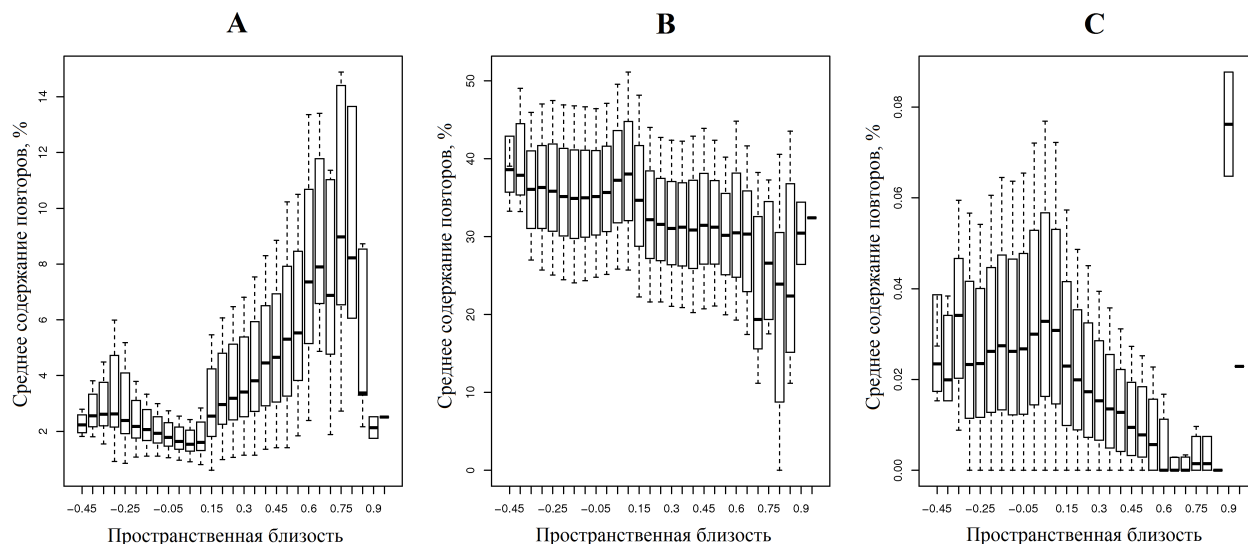


Рис. 3.6. Среднее содержание повторов в 29 рассматриваемых интервалах пространственной близости в полногеномной корреляционной матрице *C*. (А) Простые повторы. (В) Вложенные ('nested') повторы. (С) Консервативные повторы, возникшие при участии мобильных элементов ('exapted repeats').

После того, как мы завершили данное исследование, была опубликована ещё одна работа, описывающая систематические ошибки метода Hi-C [82], в которой авторы указывают на такие источники систематических ошибок, как расстояние между сайтами рестрикции, нуклеотидный состав продуктов лигирования и коэффициент выравнивания прочтений. Чтобы скорректировать эти

ошибки, авторы разработали алгоритм нормализации данных. Следуя этому алгоритму в точности как описано в [82], мы повторили их процедуру нормализации и получили полногеномную нормализованную матрицу D для лимфобластоидной клеточной линии GM06990. Мы повторили наши исследования (см. ниже) для этой матрицы D и для двух матриц из работы [40] (матрица обогащения контактами M^* и корреляционная матрица C , см. Материалы и методы) независимо, чтобы продемонстрировать, что наши результаты не зависят принципиально от алгоритма нормализации или процедуры удаления систематических ошибок.

3.2.3. Транс-сплайсинг между пространственно сближенными генами

Мы осуществили поиск химерных РНК в трех образцах по секвенированию транскриптома человека – мозговой ткани и клеточных линиях GM12878 и K562 (см. Материалы и методы). Найденные химерные РНК были наложены на матрицу пространственной близости фрагментов ДНК, чтобы проверить гипотезу о том, что пространственно близкие фрагменты ДНК образуют больше химерных РНК. Были рассмотрены только 18 центральных интервалов пространственной близости (см. выше). Для каждого интервала мы вычислили долю взаимодействующих фрагментов ДНК, между которыми наблюдается образование химерных РНК. Чтобы сделать разные наборы данных сравнимыми, это значение было в дальнейшем разделено на общее количество химерных РНК в образце. Контрольные наборы данных для каждого из трех образцов были получены путем составления искусственных пар прочтений со случайными прочтениями на другой хромосоме (см. Материалы и методы).

Данные для каждого из трех образцов показали значимые корреляции между частотой образования химерных РНК и пространственной близостью

фрагментов (коэффициент корреляции Спирмена = 0.88, 0.94, 0.85, Р-значение $< 2.2e^{-16}$, $1.7e^{-6}$, $2.2e^{-16}$, соответственно), по сравнению с контрольными наборами данных (Рисунок 3.7А). Наблюдаемая слабая корреляция между про-

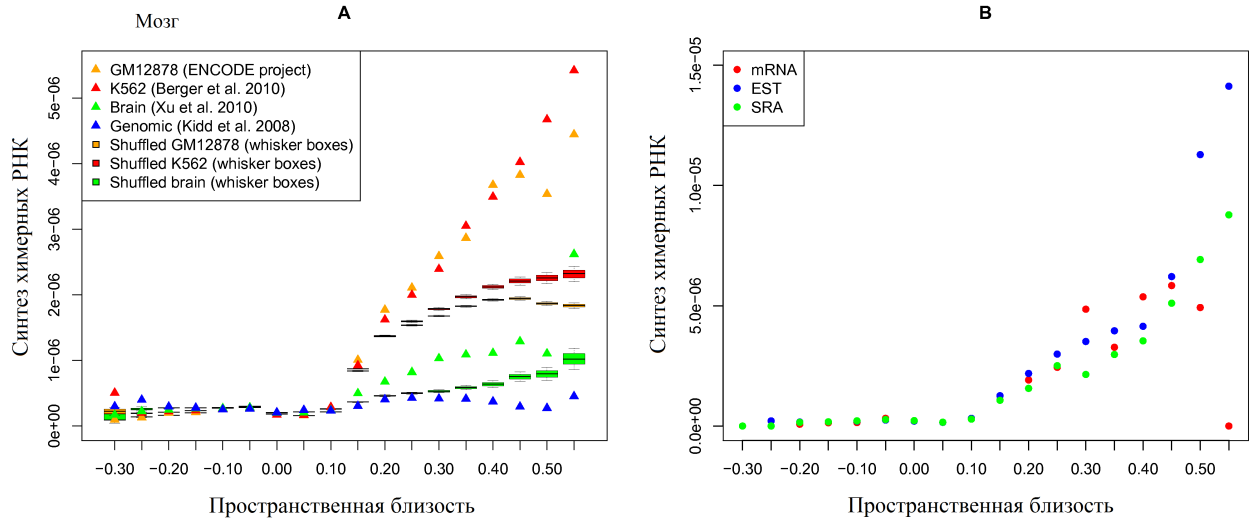


Рис. 3.7. Корреляция между производством химерных РНК и значениями пространственной близости для (А) клеточной линии K562, клеточной линии GM12878 и ткани мозга (красные, оранжевые и зеленые треугольники, соответственно); набора данных о геномных перестройках (показан синим); контрольных наборов данных K562, GM12878 и ткани мозга (красные, оранжевые и зеленые прямоугольники) и (В) трех наборов данных из базы ChimerDB: мРНК, EST и SRA (красные, синие и зеленые точки, соответственно).

изводством химерных РНК в мозговой ткани и пространственной близостью в лимфобластоидной клеточной линии особенно интересна, поскольку существование такой корреляции позволяет предположить, что трехмерная архитектура генома человека сходна, по крайней мере частично, в таких разных клетках, как лимфобласты и нейроны.

Наблюдаемые корреляции во всех наборах данных могут быть вызваны, по крайней мере, двумя причинами: транс-сплайсингом и геномными перестановками. Чтобы выяснить, какая из причин оказывает наибольшее влияние, мы проанализировали данные о геномных перестройках из работы [83] и обнаружили, что повышения доли химерных РНК среди пространственно близких

пар фрагментов в данном случае не наблюдается (Fig. 3.7A). Таким образом, найденные химерные транскрипты образуются, скорее всего, в результате транссплайсинга.

Мы также проанализировали данные из базы химерных РНК ChimerDB [74], которая содержит химерные транскрипты, собранные из различных публично доступных ресурсов. Рис. 3.7B показывает, что значения пространственной близости коррелируют с уровнем производства химерных РНК согласно всем трем источникам данных (экспрессированным коротким последовательностям, или EST, данным секвенирования, или SRA, и мРНК), доступным в ChimerDB (коэффициент корреляции Спирмена = 0.93, 0.97, 0.70, Р-значение $< 2.2e^{-16}$, $7.8e^{-12}$, 0.001, соответственно).

3.2.4. Подобие эпигенетических маркеров в пространственно сближенных участках ДНК

Мы проверили, обладают ли пространственно близкие участки генома схожим уровнем модификаций гистонов, метилирования ДНК, чувствительности к ДНКазе и экспрессии. Для этого были использованы данные нескольких исследований (см. Материалы и методы). Все перечисленные свойства имеют одну и ту же структуру данных в виде маркеров, расположенных вдоль генома. После измерения каждый маркер характеризуется пиком определенной ширины и высоты, называемым сигналом. Чтобы усреднить силу сигнала $S(i)$ на фрагменте генома i длиной в 1Мб, мы умножили высоту каждого пика (H_k) на долю (W_k) фрагмента i , пересекающегося с данным пиком, а затем просуммировали результаты для всех пиков $1...n$ во фрагменте i :

$$S(i) = \sum_{k=1}^n (H_k \cdot W_k) \quad (3.1)$$

Хотя некоторые эпигенетические маркеры точечные (например, H3K4me3 в области промотора), а некоторые занимают довольно обширные области (напри-

мер, H3K36me3 внутри гена), их тонкая структура вряд ли оказывает влияние на значение силы сигнала $S(i)$ при грубом разрешении в 1Мб. Разница $D(i, j)$ между силой сигнала в двух взаимодействующих фрагментах i и j была вычислена как:

$$D(i, j) = \left| \log \frac{S(i)}{S(j)} \right| \quad (3.2)$$

Для каждого из рассматриваемых интервалов пространственной близости в корреляционной матрице C была вычислена медиана $D(i, j)$. Оказалось, что эти медианы коррелируют со значениями пространственной близости в каждом типе данных, который был протестирован: уровень экспрессии (Рисунок 3.8А, коэффициент корреляции Спирмена = -0.96 , Р-значение = $6.0e^{-6}$), модификации гистонов (Рисунок 3.8В, средний коэффициент корреляции Спирмена = -0.85 , среднее Р-значение = $4.1e^{-4}$), метилирование ДНК (Рисунок 3.8С, коэффициент корреляции Спирмена = -0.92 , Р-значение $< 2.2e^{-16}$), чувствительность к ДНКазе (Рисунок 3.8D, коэффициент корреляции Спирмена = -0.93 , Р-значение $< 2.2e^{-16}$).

Та же процедура была повторена для матриц обогащения контактами M^* и D . Значения пространственной близости были также разделены на 29 интервалов, и только 18 центральных интервалов были рассмотрены (см. выше). Обе матрицы M^* и D показали сильные корреляции со значениями пространственной близости в каждом типе данных, который был протестирован: уровень экспрессии (коэффициент корреляции Спирмена = -0.99 и -0.93 , Р-значение = $9.9e^{-6}$ и $< 2.2e^{-16}$ для M^* и D , соответственно), модификации гистонов (средний коэффициент корреляции Спирмена = -0.64 и -0.67 , среднее Р-значение = 0.09 и 0.09), метилирование ДНК (коэффициент корреляции Спирмена = -0.85 и -0.53 , Р-значение = $1.0e^{-5}$ и 0.02), чувствительность к ДНКазе (коэффициент корреляции Спирмена = -0.85 и -0.92 , Р-значение $< 2.2e^{-16}$ и $2.2e^{-16}$). Поскольку все три матрицы C , M^* и D показали приблизительно одинаковые

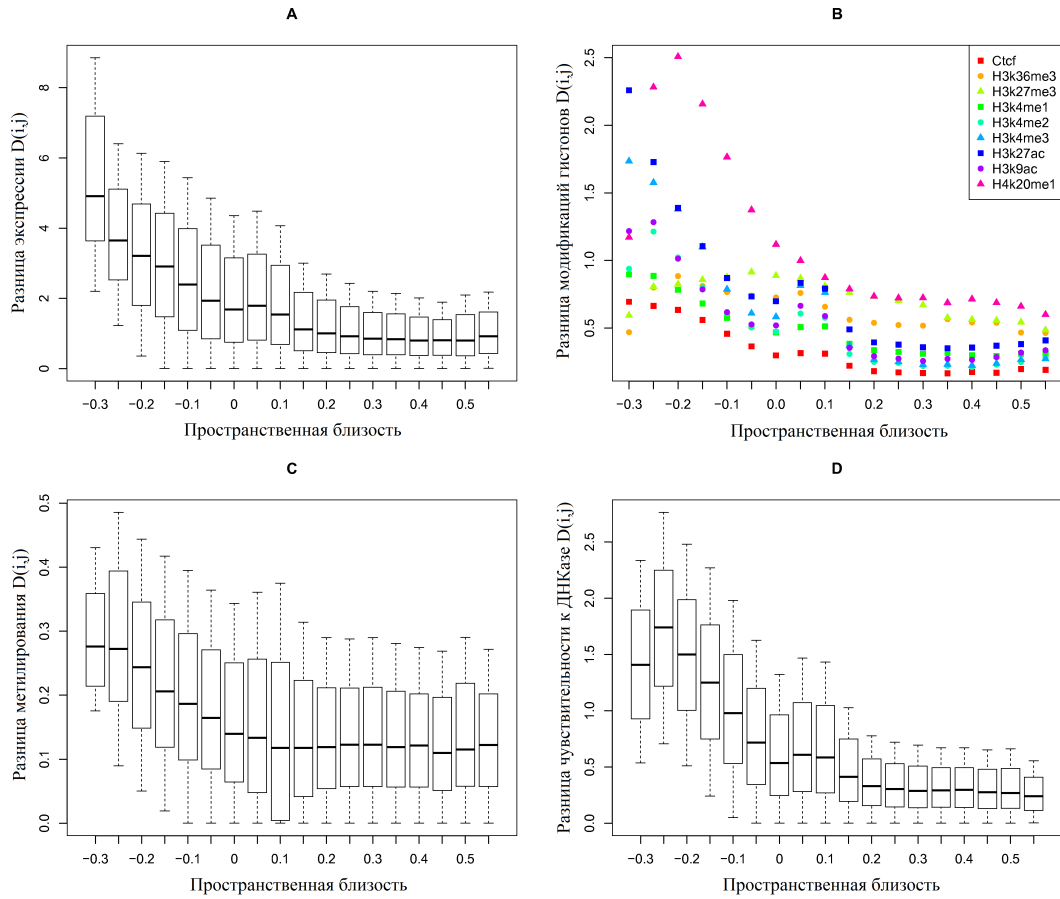


Рис. 3.8. Корреляции значений пространственной близости с разницей экспрессии (А), модификаций гистонов (В), метилирования ДНК (С) и чувствительности к ДНКазе (D). Обозначения на панелях (А,С,D) такие же, как на рис. 3.5. Символы в (В) обозначают медианы для различных модификаций гистонов.

результаты, мы выбрали только одну из них, матрицу C , для дальнейшего анализа.

Анализ контрольного набора данных

В соответствии с моделью, предложенной в работе [40], геном человека может быть условно разделен на два компартмента: компартмент А, содержащий открытый, активно транскрибирующийся хроматин, и компартмент В с противоположными характеристиками. Чтобы проверить, объясняются ли этой моделью полученные нами результаты, мы сконструировали контрольный на-

бор данных, в котором перемешали имена генов внутри каждого компармента независимо, сохраняя позиции генов нетронутыми, чтобы не разрушать участки генома, обогащенные и обедненные генами (см. Материалы и методы). Был выбран именно такой тип контроля, поскольку существует корреляция между содержанием генов и пространственной близостью фрагментов генома (коэффициент корреляции Спирмена = 0.76, Р-значение = 3.4×10^{-4}). Следовательно, возникает предположение, что наблюдаемое подобие эпигенетических маркеров в пространственно близких доменах может быть вызвано их локализацией в открытом компартменте хроматина, богатом генами. Однако, хотя наш контрольный набор данных не нарушает компартменты и плотность генов, корреляции для него значительно снижаются (средний коэффициент корреляции Спирмена = -0.67, среднее Р-значение = 0.002, сравните с -0.91 и 1.2×10^{-4} , соответственно для исходного набора данных). Это означает, что наблюдаемое подобие эпигенетических маркеров в пространственно близких фрагментах генома не может быть объяснено ни их общим происхождением из одного компартмента, ни повышенным содержанием генов. Кроме того, в контрольном наборе данных наблюдается совершенно другой корреляционный профиль: значения разницы $D(i, j)$ повышаются в несколько раз для пространственно близких фрагментов ($0.50 \leq c(i, j) \leq 0.55$) и понижаются для пространственно удаленных ($-0.30 \leq c(i, j) \leq -0.25$). Пример для разницы уровней экспрессии приведен на рис. 3.9).

Дополнительно, мы разделили пары взаимодействующих 1-Мб фрагментов на три группы: (1) оба фрагмента в закрытом компартменте; (2) оба фрагмента в открытом компарменте; (3) фрагменты в разных компартментах. Основные вычисления были повторены независимо для каждой из трех групп, и оказалось, что корреляции между подобием эпигенетических маркеров и пространственной близостью наблюдаются в каждой из трех групп и не зависят от компартмента (Рисунок 3.10).

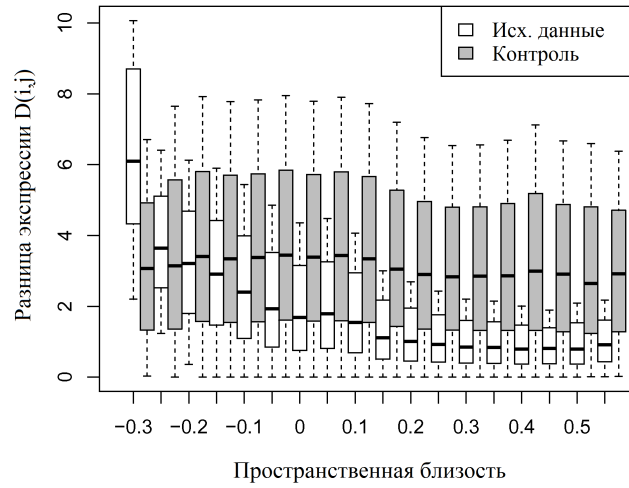


Рис. 3.9. Корреляции значений пространственной близости с разницей уровня экспрессии в исходном (показан белым цветом) и контрольном (показан серым цветом) наборах данных.

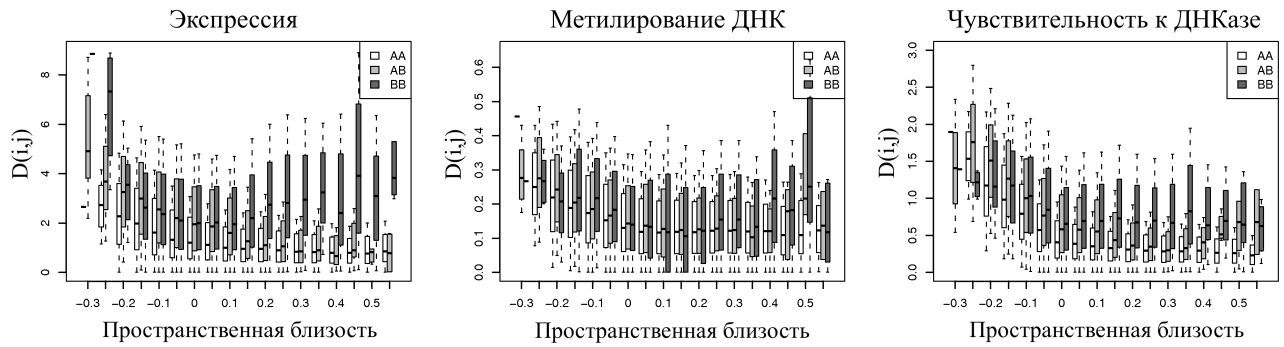


Рис. 3.10. Корреляции между подобием эпигенетических маркеров и пространственной близостью не зависят от компартмента: AA – оба фрагмента в открытом компартменте; BB – оба фрагмента в закрытом компартменте; AB – фрагменты в разных компартментах.

3.2.5. Функциональное подобие генов в пространственно сближенных участках ДНК

Далее мы проверили, распространяются ли наблюдаемые корреляции на функциональный уровень генов, согласно базе данных Gene Ontology [84]. Для этого мы определили семантическое подобие терминов (GO-терминов), описывающих функции генов. Каждому 1-Мб фрагменту генома был присвоен список GO-терминов, соответствующих генам в этом фрагменте. Чтобы вычислить среднее семантическое подобие между 1-Мб фрагментами i и j , мы составили

матрицу матрицу G с размерностью m на n для каждой пары фрагментов i и j , где m – это длина списка GO-терминов для фрагмента i , а n – длина списка GO-терминов для фрагмента j . Элементы матрицы G были вычислены с помощью программы GOSemSim [85], которая позволяет определить семантическое подобие между GO-терминами посредством алгоритма, основанного на теории графов [86]. Затем было вычислено среднее значение матрицы G .

Описанная выше процедура была повторена отдельно для трех иерархий базы данных Gene Ontology ('Молекулярная функция', 'Биологический процесс' и 'Клеточный компонент'). Оказалось, что среднее семантическое подобие GO-терминов коррелирует со значениями пространственной близости для всех трех иерархий базы данных Gene Ontology (Рисунок 3.11А-С, коэффициент корреляции Спирмена = 0.78, 0.65, 0.98, Р-значение = $2.1e^{-4}$, $4.3e^{-3}$, $8.4e^{-6}$, соответственно), с самым высоким коэффициентом корреляции для иерархии 'Клеточный компартмент'. Данное наблюдение позволяет говорить о том, что пространственно близкие участки генома обогащены генами со схожими функциями.

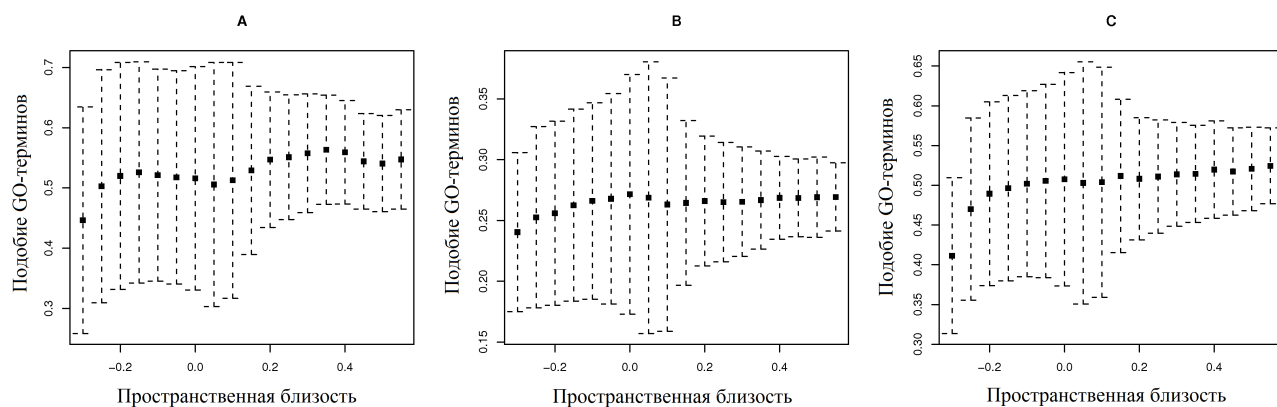


Рис. 3.11. Корреляции значений пространственной близости с семантическим подобием GO-терминов, соответствующих генам, расположенным во взаимодействующих участках генома. (А) 'Молекулярная функция'. (В) 'Биологический процесс'. (С) 'Клеточный компартмент'. Черные квадраты обозначают среднее подобие GO-терминов, пунктирные линии – стандартные отклонения.

3.2.6. Ко-экспрессия генов в пространственно сближенных участках ДНК

Мы также изучили ко-экспрессию генов в пространственно близких фрагментах. Данные были получены из базы COXPRESdb [87]. Чтобы вычислить среднюю ко-экспрессию $E(i, j)$ для двух взаимодействующих фрагментов генома i и j , мы применили формулу [3.3] для всех пар генов, $1 \dots n$ имеющих связи в COXPRESdb:

$$E(i, j) = \sum_{k=1}^n \left(\frac{W_{ki}}{N_j} + \frac{W_{kj}}{N_i} \right) \cdot R_k, \quad (3.3)$$

где W_{ki} – доля 1-Мб фрагмента i , пересекающегося с первым геном из пары связанных генов k ; W_{kj} – доля 1-Мб фрагмента j , пересекающегося со вторым геном из пары k , N_i – количество генов во фрагменте i , N_j – количество генов во фрагменте j , R_k – коэффициент корреляции Пирсона между профилями экспрессии генов в паре связанных генов k .

Для каждого рассматриваемого интервала пространственной близости были вычислены медианы значений $E(i, j)$, и для них наблюдается сильная корреляция со значениями пространственной близости (Рисунок 3.12, коэффициент корреляции Спирмена = 0.93, Р-значение $< 2.2e^{-16}$).

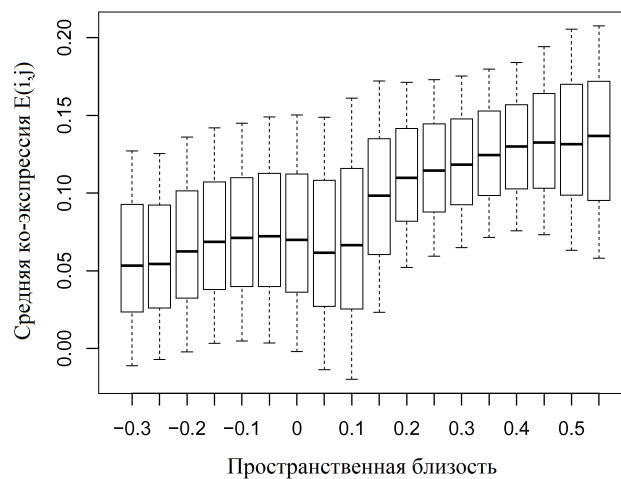


Рис. 3.12. Корреляции между значениями пространственной близости и значениями ко-экспрессии из базы данных COXPRESdb. Обозначения как на рис. 3.5.

3.2.7. Пространственно сближенные участки имеют схожие состояния хроматина

Затем мы рассмотрели так называемые состояния хроматина, биологически значимые комбинации эпигенетических маркеров [51]. Чтобы сравнить профили состояний хроматина между двумя взаимодействующими 1-Мб фрагментами генома, мы присвоили каждому 1-Мб фрагменту вектор длины n , где $n = 51$ – число состояний хроматина. Каждый элемент k данного вектора определялся как доля 1-Мб фрагмента, для которой аннотировано k -е состояние хроматина, и $k = 1..n$. Подобие двух таких векторов i и j было вычислено с помощью коэффициента Жаккарда как:

$$J(i, j) = \frac{\sum_{k=1}^n \min(i_k, j_k)}{\sum_{k=1}^n \max(i_k, j_k)} \quad (3.4)$$

Рис. 3.13 показывает, что значения пространственной близости сильно коррелируют с подобием профилей состояний хроматина (коэффициент корреляции Спирмена = 0.99, Р-значение = $9.6e^{-6}$).

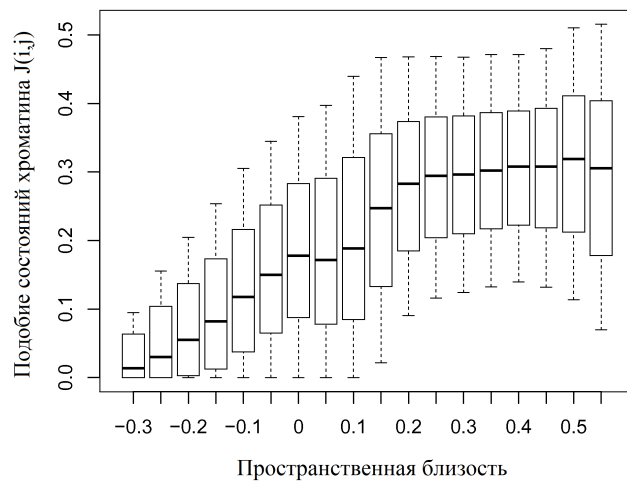


Рис. 3.13. Корреляции значений пространственной близости с подобием профилей состояний хроматина. Обозначения как на рис. 3.5.

Наблюдаемые нами корреляции между пространственным расположением

участков хроматина и подобием их состояний хорошо согласуются с теорией о транскрипционных фабриках [29]. Согласно этой теории, гены в транскрипционных фабриках образуют пространственные кластеры, обогащенные транскрипционными факторами. Эти гены характеризуются высоким уровнем экспрессии, обладают сходными эпигенетическими маркерами и выполняют похожие функции, а также часто ко-экспрессируются. Транскрипционные фабрики производят большое количество химерных РНК, в основном пост-транскрипционно, за счет транс-сплайсинга.

3.3. Выводы к Главе 3

1. Анализ систематических ошибок секвенирования показал зависимость профилей покрытия генов от лаборатории в экспериментах по секвенированию мРНК.
2. В данных о пространственной близости обнаружены и удалены систематические ошибки секвенирования.
3. Идентифицированы химерные РНК в трех наборах данных секвенирования транскриптома человека (ткань мозга, клеточная линия эритролейкемии K562, лимфобластоидная клеточная линия GM12878).
4. Пространственно близкие фрагменты ДНК образуют между собой больше химерных РНК, чем пространственно далекие, в основном за счет транс-сплайсинга.
5. Пространственно близкие фрагменты ДНК характеризуются схожими эпигенетическими маркерами и состояниями хроматина, гены в них функционально подобны и ко-экспрессируются.

Заключение

В данной работе были изучены на полногеномном уровне механизмы регуляции альтернативного сплайсинга с помощью вторичных структур РНК и белковых факторов на примере белка hnRNPL. Также было подтверждено, что транс-сплайсинг у человека может встречаться гораздо чаще, чем считалось ранее.

Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, – с помощью хэширования. С помощью этого метода у млекопитающих предсказано несколько сотен генов, содержащих консервативные вторичные структуры. Такое наблюдение позволяет предположить, что механизмы сплайсинга, опосредованные вторичной структурой, являются более распространенными, чем предполагалось ранее. Нами был также разработан метод оценки уровня ложных положительных предсказаний. С помощью этого метода было оценено, что уровень ложных положительных среди предсказанных вторичных структур не превосходит 30%. Консервативные вторичные структуры встречаются чаще ожидаемого в альтернативно сплайсируемых генах, что указывает на их участие в регуляции альтернативного сплайсинга.

Впервые получена и проанализирована полногеномная карта позиций связывания белка hnRNPL с РНК в клетках человека HeLa. Показано, что сайты связывания hnRNPL обогащены CA-повторами и CA-богатыми мотивами. Распределение позиций связывания hnRNPL вокруг 5'- и 3'-сайтов сплайсинга различается между альтернативными и константными экзонами, и между L-активируемыми и L-репрессируемыми экзонами. Это наблюдение позволяет говорить о том, что позиция связывания белка hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг. Также установлено, что белок hnRNPL часто связывается вблизи мишеней микроРНК в области 3'UTR.

Мы предполагаем, что существует глобальный механизм регуляции стабильности мРНК за счет конкуренции между микроРНК и белком hnRNPL. Также, плотность позиций связывания hnRNPL в интронах, содержащих snoRNA, значительно выше, что может говорить об участии белка hnRNPL в метаболизме snoRNA. Все эти наблюдения подтверждают гипотезу о том, что функции белка hnRNPL не ограничиваются несколькими случаями регуляции альтернативного сплайсинга. По-видимому, hnRNPL является глобальным регулятором сплайсинга, стабильности мРНК и метаболизма snoRNA.

Осуществлен поиск химерных РНК (транскриптов, предположительно образующихся в результате транс-сплайсинга) в трех наборах данных секвенирования транскриптома человека (ткань мозга, клеточная линия эритролейкемии K562, лимфобластоидная клеточная линия GM12878). Показано, что пространственно близкие фрагменты ДНК образуют между собой больше химерных РНК, чем пространственно далекие, в основном за счет транс-сплайсинга. Это наблюдение позволяет сделать вывод о более широком распространении транс-сплайсинга в клетках человека, чем считалось ранее. Кроме того, пространственно близкие фрагменты ДНК характеризуются схожими эпигенетическими маркерами и состояниями хроматина, гены в них функционально подобны и ко-экспрессируются. По-видимому, пространственно близкие фрагменты генома образуют активно экспрессирующиеся транскрипционные фабрики, что открывает новый, более сложный уровень регуляции генной активности у высших эукариот.

Выводы

1. Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, – с помощью хэширования. С его помощью у млекопитающих идентифицировано несколько сотен генов, содержащих потенциальные консервативные вторичные структуры.
2. Консервативные вторичные структуры встречаются в альтернативно сплайсируемых генах чаще ожидаемого, что указывает на их участие в регуляции альтернативного сплайсинга. В частности, показано, что образование вторичной структуры необходимо для правильного вырезания интрона в гене *SF1*.
3. Построена полногеномная карта сайтов связывания белка hnRNPL с РНК в клетках человека HeLa. Показано, что сайты связывания hnRNPL обогащены СА-повторами и СА-богатыми мотивами.
4. Распределение позиций связывания hnRNPL вокруг 5'- и 3'-сайтов сплайсинга различается между альтернативными и константными экзонами, и между L-активируемыми и L-репрессируемыми экзонами. Позиция связывания белка hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг.
5. Белок hnRNPL часто связывается вблизи мишеней микроРНК в области 3'UTR и, возможно, регулирует стабильность мРНК за счет конкуренции с микроРНК.
6. Плотность позиций связывания hnRNPL в интронах, содержащих мкРНК, значительно выше, что может говорить об участии белка hnRNPL в биосинтезе мкРНК.

7. Анализ систематических ошибок секвенирования показал зависимость профилей покрытия генов от лаборатории в экспериментах по секвенированию мРНК. В данных о пространственной близости обнаружены и удалены систематические ошибки секвенирования.
8. Идентифицированы химерные РНК в трех наборах данных секвенирования транскриптома человека (ткань мозга, клеточная линия эритролейкемии K562, лимфобластоидная клеточная линия GM12878).
9. Пространственно близкие фрагменты ДНК образуют между собой больше химерных РНК, чем пространственно далекие, в основном за счет транссплайсинга.
10. Пространственно близкие фрагменты ДНК характеризуются схожими эпигенетическими маркерами и состояниями хроматина, гены в них функционально подобны и ко-экспрессируются, что хорошо согласуется с теорией о фабриках транскрипции.

Список литературы

1. Raker V. A., Mironov A. A., Gelfand M. S., Pervouchine D. D. Modulation of alternative splicing by long-range RNA structures in *Drosophila* // [Nucleic Acids Res.](#) 2009. — Aug. Vol. 37, no. 14. P. 4533–4544.
2. Hofacker I. L., Fekete M., Stadler P. F. Secondary structure prediction for aligned RNA sequences // [J Mol Biol.](#) 2002. — Jun. Vol. 319, no. 5. P. 1059–1066.
3. Buratti E., Baralle F. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process // *Mol. Cell. Biol.* 2004. Vol. 24. P. 10505–10514.
4. Alberts B., Johnson A., Lewis J. et al. *Molecular Biology of the Cell*. New York London: GarlandScience, 2002.
5. Liu M., Grigoriev A. Protein domains correlate strongly with exons in multiple eukaryotic genomes—evidence of exon shuffling? // [Trends Genet.](#) 2004. — Sep. Vol. 20, no. 9. P. 399–403.
6. Artamonova I. I., Gelfand M. S. Comparative genomics and evolution of alternative splicing: the pessimists' science // [Chem Rev.](#) 2007. — Aug. Vol. 107, no. 8. P. 3407–3430.
7. Black D. Protein diversity from alternative splicing: a challenge for bioinformatics post-genome biology // *Cell.* 2000. Vol. 103. P. 367–370.
8. WormBook: The Online Review of *C. elegans* Biology / Ed. by T. C. elegans Research Community. Pasadena (CA): WormBook, 2005.
9. Jasinska A., Michlewski G., de Mezer M. et al. Structures of trinucleotide repeats in human transcripts and their functional implications // *Nucleic Acids Res.* 2003. — Oct. Vol. 31, no. 19. P. 5463–5468.
10. de Silanes L. I., Zhan M., Lal A. et al. Identification of a target RNA motif for RNA-binding protein HuR // *Proc. Natl. Acad. Sci. USA.* 2004. Vol. 101. P. 2987–2992.
11. Buckanovich R., Darnell R. The neuronal RNA binding protein Nova-1 recog-

- nizes specific RNA targets in vitro in vivo // *Mol. Cell. Biol.* 1997. Vol. 17. P. 3194–3201.
12. Kent O., Reayi A., Foong L. et al. Structuring of the 3' splice site by U2AF65 // *J. Biol. Chem.* 2003. Vol. 278. P. 50572–50577.
 13. Goguel V., Wang Y., Rosbash M. Short artificial hairpins sequester splicing signals inhibit yeast pre-mRNA splicing // *Mol. Cell. Biol.* 1993. Vol. 13. P. 6841–6848.
 14. Deshler J., Rossi J. Unexpected point mutations activate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron // *Genes Dev.* 1991. Vol. 5. P. 1252–1263.
 15. Chen Y., Stephan W. Compensatory evolution of a precursor messenger RNA secondary structure in the *Drosophila melanogaster* Adh gene // *Proc. Natl. Acad. Sci. USA.* 2003. Vol. 100. P. 11499–11504.
 16. Graveley B. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures // *Cell.* 2005. Vol. 123. P. 65–73.
 17. Buratti E., Muro A., Giombi M. et al. RNA folding affects the recruitment of SR proteins by mouse human polypurinic enhancer elements in the fibronectin EDA exon // *Mol. Cell. Biol.* 2004. Vol. 24. P. 1387–1400.
 18. Baraniak A., Lasda E., Wagner E., Garcia-Blanco M. A stem structure in fibroblast growth factor receptor 2 transcripts mediates cell-type-specific splicing by approximating intronic control elements // *Mol. Cell. Biol.* 2003. Vol. 23. P. 9327–9337.
 19. Graveley B. R. Sorting out the complexity of SR protein functions // *RNA.* 2000. — Sep. Vol. 6, no. 9. P. 1197–1211.
 20. Matlin A. J., Clark F., Smith C. W. Understanding alternative splicing: towards a cellular code // [Nat Rev Mol Cell Biol.](#) 2005. — May. Vol. 6, no. 5. P. 386–398.
 21. Han S. P., Tang Y. H., Smith R. Functional diversity of the hnRNPs: past,

- present and perspectives // [Biochem J](#). 2010. — Sep. Vol. 430, no. 3. P. 379–392.
22. Zuo P., Maniatis T. The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing // [Genes Dev](#). 1996. — Jun. Vol. 10, no. 11. P. 1356–1368.
 23. Lam B. J., Bakshi A., Ekinici F. Y. et al. Enhancer-dependent 5'-splice site control of fruitless pre-mRNA splicing // [J Biol Chem](#). 2003. — Jun. Vol. 278, no. 25. P. 22740–22747.
 24. Shen H., Kan J. L., Green M. R. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly // [Mol Cell](#). 2004. — Feb. Vol. 13, no. 3. P. 367–376.
 25. Hui J., Stangl K., Lane W. S., Bindereif A. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats // [Nat Struct Biol](#). 2003. — Jan. Vol. 10, no. 1. P. 33–37.
 26. Melton A. A., Jackson J., Wang J., Lynch K. W. Combinatorial control of signal-induced exon repression by hnRNP L and PSF // [Mol Cell Biol](#). 2007. — Oct. Vol. 27, no. 19. P. 6972–6984.
 27. Motta-Mena L. B., Heyd F., Lynch K. W. Context-dependent regulatory mechanism of the splicing factor hnRNP L // [Mol Cell](#). 2010. — Jan. Vol. 37, no. 2. P. 223–234.
 28. König J., Zarnack K., Rot G. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution // [Nat Struct Mol Biol](#). 2010. — Jul. Vol. 17, no. 7. P. 909–915.
 29. Gingeras T. R. Implications of chimaeric non-co-linear transcripts // [Nature](#). 2009. — Sep. Vol. 461, no. 7261. P. 206–211.
 30. Günzl A. The pre-mRNA splicing machinery of trypanosomes: complex or simplified? // [Eukaryot Cell](#). 2010. — Aug. Vol. 9, no. 8. P. 1159–1170.
 31. Li H., Wang J., Mor G., Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells // [Science](#). 2008. — Sep. Vol. 321, no. 5894.

- P. 1357–1361.
32. Rickman D. S., Pflueger D., Moss B. et al. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer // [Cancer Res.](#) 2009. — Apr. Vol. 69, no. 7. P. 2734–2738.
 33. McManus C. J., Duff M. O., Eipper-Mains J., Graveley B. R. Global analysis of trans-splicing in *Drosophila* // [Proc Natl Acad Sci U S A.](#) 2010. — Jul. Vol. 107, no. 29. P. 12975–12979.
 34. Maher C. A., Palanisamy N., Brenner J. C. et al. Chimeric transcript discovery by paired-end transcriptome sequencing // [Proc Natl Acad Sci U S A.](#) 2009. — Jul. Vol. 106, no. 30. P. 12353–12358.
 35. Dohm J. C., Lottaz C., Borodina T., Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing // [Nucleic Acids Res.](#) 2008. — Sep. Vol. 36, no. 16.
 36. Hansen K. D., Brenner S. E., Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming // [Nucleic Acids Res.](#) 2010. — Jul. Vol. 38, no. 12.
 37. Li J., Jiang H., Wong W. H. Modeling non-uniformity in short-read rates in RNA-Seq data // [Genome Biol.](#) 2010. Vol. 11, no. 5.
 38. Heap G. A., Yang J. H., Downes K. et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing // [Hum Mol Genet.](#) 2010. — Jan. Vol. 19, no. 1. P. 122–134.
 39. Schwartz S., Oren R., Ast G. Detection and removal of biases in the analysis of next-generation sequencing reads // [PLoS One.](#) 2011. Vol. 6, no. 1.
 40. Lieberman-Aiden E., van Berkum N. L., Williams L. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome // [Science.](#) 2009. — Oct. Vol. 326, no. 5950. P. 289–293.
 41. Cremer T., Cremer C., Baumann H. et al. Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromo-

- some condensation and laser-UV-microbeam experiments // *Hum Genet.* 1982. Vol. 60, no. 1. P. 46–56.
42. Hübner M. R., Spector D. L. Chromatin dynamics // *Annu Rev Biophys.* 2010. — Jun. Vol. 39. P. 471–489.
 43. Dekker J., Rippe K., Dekker M., Kleckner N. Capturing chromosome conformation // *Science*. 2002. — Feb. Vol. 295, no. 5558. P. 1306–1311.
 44. Bernstein B. E., Meissner A., Lander E. S. The mammalian epigenome // *Cell*. 2007. — Feb. Vol. 128, no. 4. P. 669–681.
 45. Kouzarides T. Chromatin modifications and their function // *Cell*. 2007. — Feb. Vol. 128, no. 4. P. 693–705.
 46. Schones D. E., Zhao K. Genome-wide approaches to studying chromatin modifications // *Nat Rev Genet.* 2008. — Mar. Vol. 9, no. 3. P. 179–191.
 47. Bock C., Lengauer T. Computational epigenetics // *Bioinformatics*. 2008. — Jan. Vol. 24, no. 1. P. 1–10.
 48. Yu H., Zhu S., Zhou B. et al. Inferring causal relationships among different histone modifications and gene expression // *Genome Res.* 2008. — Aug. Vol. 18, no. 8. P. 1314–1324.
 49. modENCODE Consortium, Roy S., Ernst J. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE // *Science*. 2010. — Dec. Vol. 330, no. 6012. P. 1787–1797.
 50. Gerstein M. B., Lu Z. J., Van Nostrand E. L. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project // *Science*. 2010. — Dec. Vol. 330, no. 6012. P. 1775–1787.
 51. Ernst J., Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome // *Nat Biotechnol.* 2010. — Aug. Vol. 28, no. 8. P. 817–825.
 52. Pervouchine D. D., Khrameeva E. E., Pichugina M. Y. et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA

- structures // [RNA](#). 2012. — Jan. Vol. 18, no. 1. P. 1–15.
53. Babak T., Blencowe B. J., Hughes T. R. Considerations in the identification of functional RNA structural elements in genomic alignments // [BMC Bioinformatics](#). 2007. Vol. 8. P. 33–33.
 54. Edgar R. MUSCLE: multiple sequence alignment with high accuracy high throughput // *Nucleic Acids Res.* 2004. Vol. 32. P. 1792–97.
 55. Solnick D., Lee S. Amount of RNA secondary structure required to induce an alternative splice // *Mol. Cell. Biol.* 1987. Vol. 7. P. 3194–3198.
 56. Hefferon T., Groman J., Yurk C., Cutting G. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing // *Proc. Natl. Acad. Sci. USA.* 2004. Vol. 101. P. 3504–3509.
 57. Mathews D., Sabina J., Zucker M., Turner H. Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure // *J. Mol. Biol.* 1999. Vol. 288. P. 911–940.
 58. Sirand-Pugnet P., Durosay P., d’Orval B. C. et al. beta-Tropomyosin pre-mRNA folding around a muscle-specific exon interferes with several steps of spliceosome assembly // *J. Mol. Biol.* 1995. Vol. 251. P. 591–602.
 59. Matsuo M., Nishio H., Kitoh Y. et al. Partial deletion of a dystrophin gene leads to exon skipping to loss of an intra-exon hairpin structure from the predicted mRNA precursor // *Biochem. Biophys. Res. Commun.* 1992. Vol. 182. P. 495–500.
 60. Huang L., Kirschke C. P., Zhang Y., Yu Y. Y. The ZIP7 gene (Slc39a7) encodes a zinc transporter involved in zinc homeostasis of the Golgi apparatus // [J Biol Chem](#). 2005. — Apr. Vol. 280, no. 15. P. 15456–15463.
 61. Langmead B., Trapnell C., Pop M., Salzberg S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // [Genome Biol](#). 2009. Vol. 10, no. 3.

62. Hui J., Hung L. H., Heiner M. et al. Intronic CA-repeat and CA-rich elements: a new class of regulators of mammalian alternative splicing // [EMBO J.](#) 2005. — Jun. Vol. 24, no. 11. P. 1988–1998.
63. Hung L. H., Heiner M., Hui J. et al. Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis // [RNA.](#) 2008. — Feb. Vol. 14, no. 2. P. 284–296.
64. Wang Z., Kayikci M., Briese M. et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions // [PLoS Biol.](#) 2010. Vol. 8, no. 10.
65. Heiner M., Hui J., Schreiner S. et al. HnRNP L-mediated regulation of mammalian alternative splicing by interference with splice site recognition // [RNA Biol.](#) 2010. — Jan-Feb. Vol. 7, no. 1. P. 56–64.
66. Jafarifar F., Yao P., Eswarappa S. M., Fox P. L. Repression of VEGFA by CA-rich element-binding microRNAs is modulated by hnRNP L // [EMBO J.](#) 2011. — Apr. Vol. 30, no. 7. P. 1324–1334.
67. Lewis B. P., Burge C. B., Bartel D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets // [Cell.](#) 2005. — Jan. Vol. 120, no. 1. P. 15–20.
68. Rossbach O., Hung L. H., Khrameeva E. et al. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L // [RNA Biol.](#) 2014. — Feb. Vol. 11, no. 2. P. 146–155.
69. Sokal R. R., Michener C. D. A statistical method for evaluating systematic relationships // [University of Kansas Scientific Bulletin.](#) 1958. Vol. 28. P. 1409–1438.
70. R Development Core Team. R: A language and environment for statistical computing / R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2009. P. 409. ISBN 3-900051-07-0.
71. Xu A. G., He L., Li Z. et al. Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq // [PLoS Comput Biol.](#) 2010. Vol. 6.

72. Berger M. F., Levin J. Z., Vijayendran K. et al. Integrative analysis of the melanoma transcriptome // [Genome Res.](#) 2010. — Apr. Vol. 20, no. 4. P. 413–427.
73. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project // [Science](#). 2004. — Oct. Vol. 306, no. 5696. P. 636–640.
74. Kim P., Yoon S., Kim N. et al. ChimerDB 2.0—a knowledgebase for fusion genes updated // [Nucleic Acids Res.](#) 2010. — Jan. Vol. 38, no. Database issue. P. 81–85.
75. Li R., Li Y., Kristiansen K., Wang J. SOAP: short oligonucleotide alignment program // [Bioinformatics](#). 2008. — Mar. Vol. 24, no. 5. P. 713–714.
76. Bohnert R., Räscht G. rQuant.web: a tool for RNA-Seq-based transcript quantitation // [Nucleic Acids Res.](#) 2010. — Jul. Vol. 38, no. Web Server issue. P. 348–351.
77. Roberts A., Trapnell C., Donaghey J. et al. Improving RNA-Seq expression estimates by correcting for fragment bias // [Genome Biol.](#) 2011. — Mar. Vol. 12, no. 3.
78. Cocquet J., Chong A., Zhang G., Veitia R. A. Reverse transcriptase template switching and false alternative transcripts // [Genomics](#). 2006. — Jul. Vol. 88, no. 1. P. 127–131.
79. Camacho C., Coulouris G., Avagyan V. et al. BLAST+: architecture and applications // [BMC Bioinformatics](#). 2009. Vol. 10. P. 421–421.
80. Karolchik D., Baertsch R., Diekhans M. et al. The UCSC Genome Browser Database // [Nucleic Acids Res.](#) 2003. — Jan. Vol. 31, no. 1. P. 51–54.
81. Lowe C. B., Bejerano G., Haussler D. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes // [Proc Natl Acad Sci U S A](#). 2007. — May. Vol. 104, no. 19. P. 8005–8010.
82. Yaffe E., Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture // [Nat Genet.](#) 2011. — Nov. Vol. 43, no. 11. P. 1059–1065.

83. Kidd J. M., Cooper G. M., Donahue W. F. et al. Mapping and sequencing of structural variation from eight human genomes // [Nature](#). 2008. — May. Vol. 453, no. 7191. P. 56–64.
84. Ashburner M., Ball C. A., Blake J. A. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium // [Nat Genet](#). 2000. — May. Vol. 25, no. 1. P. 25–29.
85. Yu G., Li F., Qin Y. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products // [Bioinformatics](#). 2010. — Apr. Vol. 26, no. 7. P. 976–978.
86. Wang J. Z., Du Z., Payattakool R. et al. A new method to measure the semantic similarity of GO terms // [Bioinformatics](#). 2007. — May. Vol. 23, no. 10. P. 1274–1281.
87. Obayashi T., Kinoshita K. COXPRESdb: a database to compare gene coexpression in seven model animals // [Nucleic Acids Res](#). 2011. — Jan. Vol. 39, no. Database issue. P. 1016–1022.